DOCUMENT RESUME

ED 385 552 TM 023 968

AUTHOR	Emmerich, Walter; And Others
TITLE	The Development, Investigation, and Evaluation of New Item Types for the GRE Analytical Measure. GRE Board Professional Report No. 87-09P.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-91-16
PUB DATE	Aug 91
NOTE	115p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Construct Validity; Correlation; Evaluation Methods; Sex Differences; Test Construction; *Test Items; *Thinking Skills; *Verbal Tests
IDENTIFIERS	*Analytical Tests; Confirmatory Factor Analysis; Exploratory Factor Analysis; *Graduate Record Examinations; Item Wording; Item Writing

ABSTRACT

The aim of this research was to identify, develop, and evaluate empirically new reasoning item types that might be used to broaden the analytical measure of the Graduate Record Examinations (GRE) General Test and to strengthen its construct validity. Six item types were selected for empirical evaluation, including the two currently used in the GRE analytical measure. Two experimental batteries, one using a three-option format and the other, a multiple yes-no format, were administered to 2 samples of approximately 370 examinees each. Item analyses and analyses of sex differences, criterion-related validity, and relationships of the experimental item types to the current GRE measures were conducted. All but one of the experimental item types exhibited promise for strengthening the GRE analytical measure, and even the one exception appeared to be a possible item type for the GRE verbal measure. Different combinations of the item types were evaluated in a series of confirmatory factor analyses, supplemented by correlational analyses and an exploratory factor analysis. The study also provided evidence that the reasoning domain consists of two major subdomains: informal reasoning and formal-deductive reasoning. Nineteen tables present analysis results. Three appendixes give examples of the experimental item types, list participating test centers, and present correlation matrices. Appendix C contains eight tables. (Contains 41 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

GRE® RESEARCH

ED 385 552

The Development, Investigation, and Evaluation of New Item Types for the GRE Analytical Measure

Walter Emmerich
Mary K. Enright
Donald A. Rock
Carol Tucker

August 1991

GRE Board Professional Report No. 87-09P
ETS Research Report 91-16

BEST COPY AVAILABLE

Educational Testing Service, Princeton, New Jersey

TM 023968

The Development, Investigation, and Evaluation of New Item Types for the GRE Analytical Measure

Walter Emmerich
Mary K. Enright
Donald A. Rock
Carol Tucker

GRE Board Report No. 87-09P

August 1991

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Graduate Record Examinations and Educational Testing Service are U.S. registered trademarks of Educational Testing Service; GRE, ETS, and the ETS logo design are registered in the U.S.A. and in many other countries.

Copyright © 1991 by Educational Testing Service. All rights reserved.

Abstract

The aim of this research was to identify, develop, and evaluate empirically new reasoning item types that might be used to broaden the analytical measure of the GRE General Test and to strengthen its construct validity. Item types were identified that varied in the aspects of reasoning they measure. Six item types were selected for empirical evaluation, including the two currently used in the GRE analytical measure. All of the experimental item types were developed in a 3-option multiple-choice format, and four of them also were developed in a multiple-yes/no format. Two experimental batteries were assembled, one using the 3-option format and the other using the multiple-yes/no format. Two samples of approximately 370 examinees each, all of whom had recently taken the GRE General Test, were administered one or the other experimental battery. Item analyses and analyses of sex differences, criterion-related validity, and relationships of the experimental item types to the current GRE measures were conducted. All but one of the experimental item types exhibited promise for strengthening the GRE analytical measure, and even the one exception appeared to be a possible item type for the GRE verbal measure. Evidence for interactions between item type and item format suggested that varying the format may result in the assessment of a different aspect of reasoning for some but not all reasoning item types. Different combinations of the experimental item types were evaluated in a series of confirmatory factor analyses, supplemented by correlational analyses and an exploratory factor analysis. Findings indicated that the convergent validity of the GRE analytical measure probably can be strengthened by selectively adding or substituting some of the experimental item types. However, such alterations of the GRE analytical measure probably would not improve the measure's discriminant validity. The study also provided evidence suggesting that the reasoning domain consists of two major subdomains: informal reasoning and formal-deductive reasoning. This outcome has implications both for understanding the structure of the reasoning domain and for predicting the impact of different combinations of reasoning item types on the construct validity of the GRE analytical measure.

Acknowledgments

This study was supported by the Graduate Record Examinations Board. We wish to express our gratitude to those who assisted in the study. Special thanks go to Clark Chalifour, Peter Cooper, Robin Huffman, Miles McPeek, and Erich Woisetschlaeger, for their fine work on developing the experimental item types and/or in writing and reviewing the experimental items. Lynn Patterson's administrative assistance was much appreciated, and Denise Wooten's assistance in arranging the test administration was invaluable. Annette Turner, the data analyst for the study, was always responsive to requests for still more analyses. We also wish to thank Neal Kingston, Miles McPeek, Don Powers, Lawrence Stricker, and Cheryl Wild for their helpful comments on an earlier version of this report.

Contents

Chapter 1. The Selection and Development of Types of Reasoning Items for the GRE Analytical Measure.....1

Chapter 2. The Empirical Evaluation of Types of Reasoning Items: Some Basic Properties.....13

Chapter 3. Improving the Convergent and Discriminant Validity of the GRE Analytical Measure.....35

Chapter 4. Summary and Conclusions.....63

References.....67

Appendix A: Examples of Experimental Item Types

Appendix B: Participating Test Centers

Appendix C: Correlation Matrices

Chapter 1

The Selection and Development of Types of Reasoning Items for the GRE Analytical Measure

In 1977 the GRE General Test was modified by the addition of an analytical measure. This measure was introduced in order to expand the definition of academic talent to include more aspects of reasoning than those included in the existing verbal and quantitative measures. Evidence was found that the new analytical measure could be differentiated from the existing verbal and quantitative measures (Powers & Swinton, 1981), and that it predicted performance in graduate school (Wilson, 1982). However, subsequent research identified problems with some of the item types included in the analytical measure. Originally, the analytical measure included four types of items. Two of these item types were found to be affected both by special test preparation (Kingston & Dorans, 1983; Powers & Swinton, 1984; Swinton & Powers, 1983) and by within-test practice effects (Swinton, Wild, & Wallmark, 1982). As a result, the two suspect item types were eliminated from the analytical measure in 1981. However, this reduction in the number of item types gave rise to other problems regarding content representativeness and construct validity (McPeek, Chalifour, & Tucker, 1985; Ward, Emmerich, Enright, Wightman, Powers, Gitomer, & Swinton, 1986). The goal of the current research was to identify, develop, and evaluate new item types that might strengthen the current GRE analytical measure.

Content Representativeness

The original GRE analytical measure did a reasonable job of representing a variety of reasoning skills, but the elimination of two of the four item types resulted in a narrower measure, one that heavily emphasizes deductive reasoning skills.

Specifically, the current version includes the logical reasoning item type (LR) and the analytical reasoning item type (AR). Logical reasoning presents a short passage in the form of an argument, usually followed by a single question. The question assesses any of a variety of critical reasoning skills, such as recognition of assumptions, evaluation of arguments and counterarguments, and analysis of evidence. The analytical reasoning item type presents a brief scenario together with a group of rules (conditions), followed by a set of four to seven questions. The questions, often stated in conditional (if-then) form, assess examinees' skills in grasping and combining rules to arrive at deductions as to what must be true or could be true, given the stated conditions. Thus, the logical reasoning item type draws on a broader range of reasoning skills than does the analytical reasoning item type, which primarily involves deductive reasoning. Because the ratio of analytical reasoning items to logical reasoning items is about three to one in the current analytical measure, the measure emphasizes deductive reasoning skills.

Convergent and Discriminant Validity

A test's internal structure should be consistent with the conceptualization of the skills the test is designed to measure. Specifically, item types that are supposed to assess the same or highly similar skills should be more highly correlated with one another than with item types designed to measure different skills. Such is the case for the item types in the verbal and quantitative measures of the GRE General Test. For example, Wilson (1985) reported correlations among the verbal item types ranging from .61 to .84 and correlations among the quantitative item types ranging from .64 to .71, supporting the convergent validities of the two measures. The correlations of the verbal item types with the quantitative item types were considerably lower, ranging from .34 to .49, supporting the discriminant validities of the two measures. However, such a pattern of relationships has not been found for the GRE analytical item types currently in use. Generally, logical reasoning correlates more highly with the verbal item types than it does with analytical reasoning, and analytical reasoning correlates more highly with the quantitative item types than with logical reasoning (e.g., Wilson, 1985). This situation constitutes a threat to the coherence of the analytical measure because the reported scores on that measure are derived by combining the scores on logical reasoning with those on analytical reasoning.

This problem has a direct impact on the test development process. During pretesting, the quality of each pretested item is determined in part by correlating the item with a criterion measure based on the total analytical score, consisting of the sum of the scores on both the logical reasoning items and the analytical reasoning items. Thus, because the criterion (total analytical score) consists of two item types that correlate better with the verbal or with the quantitative item types than they do with each other, item attrition is quite high during pretesting Moreover, because the GRE analytical measure is composed of more analytical reasoning items (76%) than logical reasoning items (24%), there exists an imbalance in the criterion measure that probably contributes to the far greater attrition rate of logical reasoning items during pretesting.

Improving the Analytical Measure: Research Plan

The goal of this study was to develop and evaluate experimental item types that might help strengthen the existing GRE analytical measure. The aim was to broaden the measure and to enhance its convergent and/or discriminant validity, but without adding new dimensionality and without focusing on possible alterations of the verbal measure and/or of the quantitative measure of the current GRE General Test (Ward et al., 1986).

The research was carried out in three phases. First, the new item types were to be identified and developed, as described in Chapter 1. As part of this effort, examples of each item type were to be written and reviewed to evaluate whether the item type could yield sound items capable of contributing to an improved analytical measure. In addition, the effect of modifying the current 5-option multiple-choice format was also considered. The outcome of this first phase was the development of experimental batteries of items that

included examples of four new item types as well as examples of the two current item types, but in modified formats. In the second phase, described in Chapter 2, the new item types were evaluated empirically by including them in experimental batteries administered to examinees who had recently taken the GRE General Test. These items were then evaluated in terms of their psychometric characteristics, their relationships to gender, their concurrent validities, their relationships to one another, and their relationships to the GRE verbal and quantitative measures. In the final phase, described in Chapter 3, factor analyses were conducted to assess the potential of the experimental item types for improving the construct validity of the analytical measure.

Guidelines for the Selection of New Item Types

In the first phase of the study, four new item types were selected for further evaluation. In addition, format variations were seen to offer a number of potential benefits. We now turn to a consideration of the factors that guided the selection of the new item types, descriptions of the new item types, and the rationale for exploring format variations.

Initially, members of the Educational Testing Service (ETS) test development and research staffs reviewed a number of reasoning item types. In these reviews, the following considerations were of primary importance:

1. Recent discussions of reasoning skills relevant for the analytical measure were to be taken into consideration in selecting promising item types (e.g., Powers & Enright, 1987; Tucker, 1985).

2. New dimensionality was not to be introduced into the analytical measure because the addition of another dimension would be likely to undermine convergent validity. This meant, for example, that figural stimulus materials and/or response options could not be introduced.

3. A reasoning item type that is suitable for the General Test may, nonetheless, have domain-specific content. For example, one proposed item type, a revised version of analysis of explanations, appeared to be somewhat more suitable for content from the sciences or the social sciences than for content from the humanities. To achieve balance in this regard, an effort was made to include at least one item type that would be especially suitable for content from the humanities.

4. Beyond the needs for unidimensionality and content balancing, perhaps the most severe constraint, at least on improving the analytical measure's discriminant validity, arose from the very nature of the GRE General Test as a whole. Specifically, for the General Test, the analytical measure does not uniquely measure reasoning skills because both the verbal and quantitative measures of the General Test also measure reasoning skills. For this reason, and given other constraints, it would have been unrealistic to expect that the analytical measure could be made as distinct from the verbal and quantitative measures as these measures are from each other (McPeek et al., 1985). At best, a revised analytical measure might assess reasoning skills more thoroughly than does either the verbal measure or the quantitative

measure. Earlier evidence had indicated that this rather modest goal was realistic. The earlier version of the analytical measure, consisting of four item types rather than two, had better discriminant and convergent validity (Powers & Swinton, 1981) than does the current measure.

Description of the New Item Types

Our review and selection of promising new item types was facilitated by an examination of a compendium of reasoning item types prepared by Carlton (1987). In addition, two item types--analysis of explanations and matrix completion--were targeted for special consideration (Ward et al., 1986). On the basis of this initial review, four item types were selected for further development. These item types and the factors that entered into their selection and development are described below. Examples of each of the item types are presented in Appendix A.

Analysis of Explanations (revised). In this item type, a situation is described in a passage and a result is stated that seems paradoxical in terms of the situation and so requires explanation. The examinee is then asked to consider each of several statements. For some statements, the examinee is asked to decide whether the statement is or is not relevant to any possible adequate explanation of the result. For other statements, the examinee is asked to judge whether the statement could adequately explain the result.

Studies by Tucker (1985) and by Powers and Enright (1987) had suggested the importance of this aspect of reasoning, which was not well represented either by the logical reasoning item type or by the analytical reasoning item type. In the Powers and Enright study, this skill was called "generating valid explanations," and it was part of a broader factor "defined primarily by variables related to the drawing of conclusions" (p. 7). In the Tucker study, the skill that was rated most important with highest consistency was "formulating alternative possibilities of conceptualization, classification, or explanation" (p. 11).

In the original version of the analytical measure, analysis of explanations was presented in a fixed response format, with the five answer options being the same for all questions. Although initial studies had indicated that this item type had desirable psychometric properties (Miller & Wild, 1979), it was subsequently dropped from the General Test because it appeared to be susceptible to practice and coaching effects (Swinton & Powers, 1983; Swinton, Wild, & Wallmark, 1982). A number of factors may have contributed to these effects. Analysis of explanations had complex instructions and did not follow the "choose the best answer" convention. Learning the "tree structure" of the fixed response format may also have contributed to the practice effect. In the current project, however, this item type was revised to include options that would be unique to each item, an approach that appeared to circumvent problems associated with the fixed response format.

Numerical Logical Reasoning. This item type was based on work by Ward, Carlson, and Woisetschlaeger (1983) in their attempt to develop "ill-structured" problems in a multiple-choice format. "Well-structured" problems

are deductive in nature, requiring the manipulation of symbols as tokens and the application of algorithms. "Ill-structured" problems are more complex, do not have definite criteria for determining when a problem is solved, and lack some of the information needed to solve the problem (Simon, 1978). In the "ill-structured" problems developed by Ward et al. (1983), the stimulus material is presented in the form of a chart, graph, or table. The examinee is presented with, say, a table; the question asks the examinee to analyze or to evaluate a stated finding or other information in the table. For example, two contrasting interpretations of the material given might be presented, and the examinee asked to select the option that best supports one of those interpretations. As another example, the examinee might be asked to select the most plausible explanation for the information in the table.

This item type was of interest because it calls for an integration of numerical and verbal reasoning in a problem-solving context. This item type had been found to correlate quite highly with the logical reasoning item type and with the former analysis of explanations item type (Ward et al., 1983).

Contrasting Views. In this item type, two contrasting views are presented, followed by a series of questions bearing on both of the views. Each view centers on a term or concept that is expressed in each view, but that nevertheless has different implications within the two views. The two views can be seen as alternative interpretations of the concept. Some of the questions measure the ability to recognize common aspects (central concepts or common assumptions), whereas others focus on aspects of disagreement (differences in implications or interpretation). Still other questions measure the ability to determine the relationship of a third view to the two given views. This item type is a variant of an item type called contrasting arguments (Carlton, 1987).

This item type's emphasis on the evaluation of different but related views was attractive because of its suitability for materials from the humanities and for measuring the ability to "notice significant details or anomalies," an aspect of reasoning that was rated as being very important in graduate study (Tucker, 1985).

Pattern Identification. The analytical measure does not now include an item type designed to assess what is commonly called "inductive reasoning." For this reason, we had hoped to adapt a matrix completion item type, such as Raven's Progressive Matrices (Raven, 1965). Unfortunately, the development of a workable matrix completion item type proved to be difficult if not impossible. As noted earlier, the use of figural stimuli was precluded, and some of the most successful matrix item types in the literature, such as Raven's Progressive Matrices, rely heavily on figural materials. Also, it was necessary to provide clear (and not too lengthy) instructions to the examinees. Because matrix tasks involve two dimensions (rows and columns), and because the instructions also would have to include other constraining information (discussed later), it was concluded that an adequate explanation of the item type's two-dimensional feature would have stretched the instructional load beyond reasonable limits.

On the other hand, a series item type deals with only a single dimension, and, because matrix problems and series problems have been found to measure similar cognitive skills, attention was shifted to the development of a series item type. The item type we had in mind would require the examinee to construct an applicable sequence rule and then to apply that rule by selecting another sequence (option) based on the same sequence rule.

Several challenges were encountered along the way, however. We considered using letter series, but this approach ultimately comes down to using the letters in numerical ways, and the use of letters might be unfair to examinees who have difficulty discriminating among the letters in certain classes of letters. We also gave some thought to constructing a "word series" item type in which the successive words in a series might be synonyms, opposites, part-wholes, etc., but this idea was abandoned because the resulting item type would have been too close to the present verbal item types, perhaps further jeopardizing the analytical measure's discriminant validity.

The remaining possibility was to draw on the familiar number series item type. However, further obstacles had to be overcome before this item type could be considered as a candidate for the analytical measure. To illustrate, given an incomplete number series, such as 2, 4, 6, 8, __, it seems obvious to most people that the missing number must be 10. Indeed, past publishers of tests that include number series have presumed the truth of such a conclusion in developing their item keys. But mathematicians and logicians have observed that the intended answer is not uniquely determined. We asked a mathematician to investigate whether sequence rules other than those intended could be used to answer a few series completion items correctly. The mathematician found that alternative rules were possible, resulting in unintended multiple answers. Although we have not proven formally that any number could substitute, say, for 10 in the above example, or for any number in any series, such possibilities are to be taken seriously. The natural tendency is to dismiss this problem as trivial because very few examinees are interested in playing the mathematical game required to undermine a test maker's intent. Yet this would seem to be a fair game to play, and just one successful effort would be enough to undermine the item type's defensibility. To the best of our knowledge, this problem had never been resolved previously by test makers.

In the present formulation of this item type, a sequence of numbers is presented, and the examinee is asked to select, from a set of answer options, another sequence of numbers whose pattern matches that embodied in the first sequence. The key feature of this approach is that the examinee must construct an applicable series rule. However, in order to ensure that the correct answer is unique and defensible, constraints were placed on the rule that could govern the number sequence. For example, the permissible operations referred to in the rule are to be limited to addition, subtraction, multiplication, and division of positive integers less than or equal to 3. Because of these (and other) constraints, we have called this item type pattern identification. The instructions for this item type, which state the constraints, are given in Appendix A.

Rationale for Exploring Variations in Item Format

Each item in the current GRE analytical measure is cast in a 5-option multiple-choice format. It became apparent that the effects of reducing the number of options (per item) should be explored. Specifically, we considered a 3-option multiple-choice format and a 2-option multiple-yes/no format.

The 3-Option Multiple-Choice Format. Theoretical analyses of the effects of the number of options had indicated that a 3-option multiple-choice format might be optimal. Lord (1977, 1980) reviewed four theoretical approaches to determining the optimal number of options on multiple-choice tests. Although the four approaches reviewed differed in their definition of optimal features, they all suggested that the 3-option multiple-choice format would be optimal in most situations. For example, Lord's analyses of a test's psychometric efficiency indicated that 3-option tests are optimal for middle-level examinees and equally efficient for low- and high-level examinees. Large differences in test efficiency for low- and high-level examinees were found for tests with more or fewer options.

The results of other investigations reveal a less clear picture, however. Ruch and Stoddard (1925) found lower reliabilities for 3-option items than for items with 2 or 5 options, based on an achievement test in history and the social sciences. On the other hand, Costain (1970, 1972) reported equal or higher reliabilities for 3-option versions of course examinations in psychology than for 4-option versions. Budescu and Nevo (1985) suggested that these discrepancies arise because theoretical analyses, such as Lord's, are based on the assumption that test-taking time is proportional to the total number of options on a test. In this case it is assumed that a reduction in the number of options (per item) can be balanced by an increase in the number of items so the total number of options on the test remains constant without increasing the time necessary to take the test. Budescu and Nevo note that the validity of this assumption varies with the type of item and its processing requirements. They suggest that this assumption is more likely to be true for items with simple stems than for those with very complex stems. In their empirical study of variations in the number of options on a vocabulary test, a test of mathematical reasoning, and a reading comprehension test, these researchers failed to find support for this assumption. Although Budescu and Nevo suggest that the optimal number of options is likely to be more than three, they do not propose a solution to the problem of the optimal number of options. However, their research demonstrates the importance of empirical investigations of the topic, especially the need to take into account the processing requirements of item types.

The Multiple-Yes/No Format. In a multiple-yes/no format, the examinee must decide whether each of a number of options associated with a question is correct or not. One potential advantage of this format is an increase in the amount of information extracted from each question stem. For example, with a 3-option multiple-choice question, one bit of information is extracted: whether or not the examinee has selected the correct answer. However, if this question is modified into a multiple-yes/no format in which the examinee has to judge the correctness of each option, three bits of information (whether or not the examinee thinks each option is correct or incorrect) are obtained

without adding much time to reading the question and its options. Of course, the probability of guessing each correct response is higher with a multiple-yes/no format (.50) than with a 3-option multiple-choice format (.33). Nevertheless, the effect of this difference in guessing on test reliability might be offset by increasing the number of items (Ebel, 1969; Grier, 1975). Thus, for example, one 3-option item can be equal to three multiple-yes/no items. On the other hand, high correlations might occur among the responses to yes-no options associated with the same stem (Albanese & Sabers, 1988), perhaps resulting in dilution rather than enhancement of measurement.

The possibility of reducing the number of options and thereby the time needed by the examinee to complete each item was attractive because time is at a premium in taking the analytical measure. Consider, for example, a section of a test that now has six logical reasoning items, each with its own stimulus material (passage), and each followed by a single 5-option multiple-choice question. Suppose, instead, we were to substitute 10 items in a 3-option format, and were to include several reasoning questions per passage rather than just one. If the reading comprehension load of the analytical measure could be so reduced, without reducing its reasoning load, not only might there be a time savings, but perhaps the analytical measure's discriminant validity in relation to the verbal measure might also be enhanced.

The multiple-yes/no format appeared to be especially suitable for the revised version of analysis of explanations. When an examinee is asked to evaluate the relevance or correctness of each of several statements in relation to an argument, result, or conclusion, each of the statements can stand alone as an independent consideration, and such independence is a natural feature of many problem-solving situations. In this regard, we suspected that deciding whether a statement is applicable or correct (yes/no) differs from deciding which of several stated options is most applicable or correct (multiple-choice). Specifically, the yes/no format provides less structure: lacking knowledge that exactly one of the options must be correct, the examinee cannot employ a strategy of eliminating incorrect answers (distracters), a strategy that is useful when answering multiple-choice reasoning questions but that appears to be of limited relevance for the reasoning tasks called for by graduate study.

Of course, other construct-irrelevant strategies might be applied to the multiple-yes/no format, but they can be controlled. Tendencies to say "yes" when in doubt or "no" when in doubt can be minimized by ensuring that each section of a test has about equal numbers of correct "yes" and "no" answers, and by informing the examinees of that fact in advance, thereby tending to neutralize irrelevant response sets or styles. In sum, use of the multiple-yes/no format was thought to be yet another possible way to enhance the construct validity of the analytical measure.

Instrument Development

The analysis of explanations and numerical logical reasoning items were developed in both the 3-option multiple-choice format and in the multiple-yes/no format, whereas pattern identification and contrasting views items were

written in the 3-option multiple-choice format only. Because our highly constrained version of pattern identification was so new, we decided to try it out in the more familiar multiple-choice (3-option) format only. Because the stems appropriate to the contrasting views item type often required a "best" or "most appropriate" answer, rather than a yes/no answer, this item type also was developed in the 3-option multiple-choice format only.

Two experimental batteries of the new item types were developed, one in a 3-option multiple-choice format and one in a multiple-yes/no format. Table 1 lists the order of the blocks of various item types and the number of items per item type included in the experimental batteries. As a rule of thumb for producing satisfactory internal-consistency reliabilities for each item type, considered alone, we calculated the ratios of the probabilities of guessing the correct answer in the various formats. Specifically, we calculated the ratio of the probabilities of guessing the correct answer as 1.65 for 3-option versus 5-option items and as 2.50 for multiple-yes/no versus 5-option items. This ratio was multiplied by 10, the number of multiple-choice items assumed to be necessary to provide satisfactory reliability when each item presents five options.

Thus, we estimated that, in order to achieve satisfactory internal-consistency reliabilities, about 16-17 items of each item type would be needed in the 3-option battery, and about 25 items of each type would be needed in the multiple-yes/no battery. However, because fewer item types were included in the multiple-yes/no battery, more than the minimum number of each item type were included in this battery, in part because we also wanted to balance the two test batteries in terms of the total testing time required. The time estimated as necessary to complete each item was .8 minute for multiple-yes/no items and 1 minute for 3-option multiple-choice items. Additional testing time was allowed to facilitate mastery of the instructions for each of the two experimental item types having extensive instructions (analysis of explanations and pattern identification).

Items representative of the experimental item types were developed by experienced ETS staff in test development. To ensure item quality, we employed essentially the same item-development procedures used in developing items for the GRE General Test, including at least two independent reviews of each item. For some of the experimental item types (analytical reasoning, logical reasoning, and analysis of explanations), it was possible to draw from the pool of items previously used in the General Test and to translate the items from the 5-option format to the 3-option multiple-choice and multiple-yes/no formats. For the remaining experimental item types (contrasting views, numerical logical reasoning, and pattern identification), which had never been used previously on the GRE General Test, the staff attempted to write items that would reflect the average difficulty level and the spread of difficulty levels of items currently included in the GRE analytical measure.

When items were written in both the 3-option multiple-choice and multiple-yes/no format, the transformation from one to the other was not mechanical, even though the same basic stimulus material was used. Three-option items have one correct and two incorrect options. In the multiple-

Table 1

Order of Item Blocks and Number of Items
in Experimental Batteries

Format			
Multiple Yes/No		3-Option Multiple-Choice	
Item Block	No. of Items	Item Block	No. of Items
Section 1		Section 1	
AR	10	AR	5
LR	9	LR	4
AX	18	AX	8
NLR	4	PI	8
Section 2		Section 2	
NLR	14	CV	9
AR	10	NLR	8
LR	12	LR	4
NLR	9	AR	6
		LR	4
		NLR	6
Section 3		Section 3	
NLR	9	NLR	2
AX	18	CV	9
LR	9	PI	8
AR	10	AX	8
		LR	4
	—	AR	5
Total Items	132		98
Total for Each Item Type			
AR	30	AR	16
LR	30	LR	16
AX	36	AX	16
NLR	36	NLR	16
		PI	16
		CV	18

AR = Analytical Reasoning
LR = Logical Reasoning
AX = Analysis of Explanations
NLR = Numerical Logical Reasoning
PI = Pattern Identification
CV = Contrasting Views

yes/no format, however, an approximately equal balance of "yes" and "no" items was necessary. Therefore, if material was to be transformed from a yes/no to a 3-option format, as was done for the analysis of explanations item type, new distracters needed to be written.

When material was to be transformed from the 3-option to a yes/no format, some of the incorrect options in the 3-option format had to be dropped or changed to become correct answers. Further, not all of the problems posed in a 3-option format were retained in the yes/no format, because doing so would have generated many more yes/no items than were needed. As a result, analytical reasoning items cast in the yes/no format comprised a much-reduced exploration of the implications of the given conditions, relative to the set of questions in the 3-option format. For the logical reasoning item type, in contrast, though fewer problems were presented in the yes/no format than in the 3-option format, the depth of analysis was not markedly different in the two formats. Reviewer comments also sometimes resulted in language adjustments, especially when the existing language did not exactly fit the question to be asked.

Chapter 2

The Empirical Evaluation of Types of Reasoning Items: Some Basic Properties

A central aim of the study was to determine which combinations of item types, if any, might strengthen the current GRE analytical measure. Important information for making such judgments includes the basic characteristics of the items, described in this chapter, and their potential for improving the convergent and discriminant validities of the current analytical measure, the focus of the analyses presented in Chapter 3.

Method

Design

The experimental batteries of the new item types were administered to a sample of examinees who had all taken the same form of the GRE General Test in December 1988. This procedure enabled us to analyze relationships of the experimental item types to the item types currently used in the General Test. Different groups of examinees were tested on the two experimental batteries, which differed in format (3-option multiple-choice vs. multiple-yes/no).

Additionally, for purposes of controlling for practice effects and fatigue-boredom effects, two alternate forms of each of the two experimental batteries were constructed. The blocks of items (see Table 1) were presented in reverse order on the alternate forms of each battery.

Procedures

A special test administration for the experimental batteries was arranged for April 15, 1989, at 35 of the regular testing centers for the GRE General Test. These test centers were widely distributed throughout the United States (see Appendix B). In the fall of 1988, a small-scale pilot test had been conducted at two test centers near ETS to try out and fine-tune the administration procedures and test-taking times and to obtain an estimate of the participation rate of examinees in the fuller study.

Subject Recruitment. In January 1989, all of the 7,005 U.S. citizens who took the December 1988 GRE General Test at the selected test centers were sent letters and asked if they would be interested in returning to the same test centers in April to participate in an experimental study of new types of test questions. They were offered $35 for their participation. About 2,000, or 29%, of those contacted, indicated that they were interested in participating. Because a total sample size of about 800 (about 400 per battery) was considered adequate, and pilot testing had shown that only about 50% of those subjects who had expressed interest actually showed up, admission tickets for the special administration were sent to 1,600 of the respondents. The other 400 respondents were eliminated randomly from the subject pool, with the constraint that extra examinees first be eliminated randomly from those test centers where the seating capacity might be exceeded.

Test Administration. At each test site, the experimental batteries were administered by experienced staff who were familiar with ETS testing procedures. Regular test procedures were followed to the extent permitted by the formats of the experimental batteries.

The experimental test consisted of three 45-minute sections separated by short breaks. To compensate for the unfamiliarity of the item types and to minimize test speededness, the time allocated for each section was 5 to 10 minutes longer than the time estimated as necessary to complete the items. In addition, a 5-minute period was added at the end of the test session during which the examinees answered a short questionnaire regarding the experimental item types.

Examinees marked their answers directly in the test booklets (separate answer sheets were not used). The four test forms (multiple-yes/no version, orders 1 and 2; 3-option version, orders 1 and 2) were spiraled, with the multiple-yes/no and 3-option forms alternating.

Results

The initial analyses of the experimental data were designed to determine whether the experimental sample was representative of the population that took the GRE General Test in December 1988 as well as some of the basic properties of the experimental item types.

The Sample

Of the 1,600 examinees sent admission tickets for the experimental test, 762 appeared at the test centers and completed the experimental test. Preliminary analyses identified two examinees whose scores on the experimental test were both below chance and three standard deviations lower than what would have been expected on the basis of their scores on the GRE analytical test taken in December 1988. Because these data suggested that these two examinees did not take the experimental task seriously, their data were excluded from further analyses. Data from a few other examinees who scored at or below chance on an experimental battery were not excluded because they also had received comparatively low scores on the December 1988 analytical measure.

Table 2 presents the mean GRE scores for the December 1988 examinees (U.S. citizens only), and for the experimental sample. Results are presented separately by sex as well as for the total groups. The proportion of females (.61) in the experimental sample was slightly higher than the proportion of females (.58) among the December test takers. The ethnic composition of the sample (.86 White, .06 Black, .04 Asian American, .02 combined Hispanic groups, .005 Native American) was similar to that of GRE test takers overall, but the small size of the sample precluded analyses of ethnic differences in performance.

Table 2

Mean GRE Test Scores for December 1988 Examinees
(U.S. Citizens Only) and the Experimental Sample

Measure	December 1988 Examinees	Experimental Sample
Verbal		
Females		
Mean	502	518
SD	114	114
N	34,300	465
Males		
Mean	529	553
SD	116	110
N	24,982	295
Total		
Mean	513	532
SD	116	114
N	59,373	760
Quantitative		
Females		
Mean	506	534
SD	121	122
N	34,300	465
Males		
Mean	588	622
SD	129	116
N	24,982	295
Total		
Mean	541	568
SD	131	127
N	59,373	760
Analytical		
Females		
Mean	543	560
SD	121	125
N	34,300	465
Males		
Mean	565	593
SD	125	121
N	24,982	295
Total		
Mean	552	573
SD	123	124
N	59,373	760

As seen in Table 2, the mean GRE scores for the examinees who took the experimental test were slightly higher than those for all U.S. citizens who took the test in December 1988. Although the mean differences were highly significant, the number of cases was large, and the differences represent only .20 to .25 of the standard deviations of the test scores, which is a small difference by conventional standards (Cohen, 1977). Note also that the standard deviations for the experimental samples are about the same as for the entire group, an important outcome because we wanted the distribution of scores in the experimental sample to be as variable as it was in the larger group. Males had higher scores on all three GRE measures in both the entire group and in the experimental sample. In sum, the experimental sample had higher test scores and a higher proportion of females than the larger group, but these differences were small enough to enable us to generalize the study's findings to the candidate population.

The distribution of the four spiraled test forms at the test sites divided the total experimental sample into four smaller groups. The mean December 1988 scores for the four groups on the three GRE General Test measures are presented in Table 3. For the total groups taking the two batteries, there were no significant differences on any of the measures. However, there was a significant difference on the GRE analytical measure between examinees taking Order 1 and Order 2 on the 3-option multiple choice battery ($t = 2.05$, $df = 372$, $p < .04$). For the multiple-yes/no format, examinees taking Order 1 and Order 2 differed on the GRE verbal measure, a difference that approached significance ($t = 1.92$, $df = 384$, $p < .06$). In both of these cases, however, the score differences were quite small, representing only .20 of the standard deviation. Because these differences appeared to be of little practical importance, the subgroup data were combined, as appropriate, in the subsequent analyses.

Analyses of the Experimental Item Types

Overview

The analyses of the experimental item types were conducted with several aims in mind. First, traditional test and item analyses were performed to screen out any weak items (in terms of content and/or low discrimination power) and to determine whether the experimental item types showed acceptable levels of difficulty and reliability. Second, evidence bearing on criterion-related validity was provided by relating examinees' performance on the experimental item types with their self-reported undergraduate grades. Third, observed sex differences on the experimental item types were considered. Fourth, the experimental item types were examined for evidence of practice and fatigue-boredom effects. Fifth, correlational analyses were carried out to determine relationships of the experimental item types to one another and to measures from the GRE General Test. The results of these analyses are presented below. Finally, the findings for the experimental item types are summarized and the influences of item format are reviewed.

Table 3

Mean December 1989 GRE General Test Scores for Experimental Samples

	Sample of Examinees Administered					
	3-Option			Multiple Yes/No		
GRE Measure	Order 1 (N=196)	Order 2 (N=178)	Total (N=374)	Order 1 (N=202)	Order 2 (N=184)	Total (N=386)
Verbal						
Mean	527	542	534	540	518	528
SD	114	114	114	116	109	115
Quantitative						
Mean	562	574	568	576	562	569
SD	127	127	127	124	130	128
Analytical						
Mean	562	589	575	579	564	570
SD	126	129	128	118	124	122

Table 4 provides the abbreviations used throughout the remainder of this report to identify the various item types and their formats.

Test and Item Analyses

Speededness. As noted earlier, the time limits designated for the various sections of the experimental batteries appeared to be generous, and, in fact, 100% of the examinees attempted all of the items. Also, a majority of the examinees indicated (in the questionnaire) that they felt little or no time pressure during the experimental test session. However, examinees who took the 3-option multiple-choice battery experienced somewhat more time pressure than did those who took the multiple-yes/no format. When asked how much time pressure they experienced overall, 95% of the examinees taking the multiple-yes/no battery and 68% of the examinees taking the 3-option form responded "little or none." In response to another question, 99% of the multiple-yes/no group and 84% of the 3-option group indicated that the time pressure in taking the experimental battery was less than that in taking the General Test.

These results indicated that the item and reliability analyses of the experimental item types would not be inflated by speededness. However, caution should be exercised in drawing conclusions from the differences in reported time pressures because the present study was not designed to determine the differential amount of information gained per unit of time for items cast in the two formats.

Difficulty. Information about the difficulty, reliability, and discrimination power of the experimental item types is presented in Table 5. For purposes of comparison, Table 5 also includes similar information regarding the analytical items in the December 1988 GRE General Test (AR5 and LR5). It will be recalled that the item types included in both the multiple-yes/no and the 3-option batteries had individual items of very similar content. However, it should also be noted that the individual AR5 and LR5 items included in the December 1988 GRE General Test differed in content from all of the AR and LR items included in each of the experimental batteries.

In Table 5, difficulty is reported as mean percentage correct. Theoretically, the probability of guessing the correct answer increases as the number of options decreases. For 5-option, 3-option, and multiple-yes/no items, the probability of guessing the correct answer is .20, .33, and .50, respectively. Thus, the mean percentage correct that would be expected by chance for sets of 5-option, 3-option, and multiple-yes/no items are 20%, 33%, and 50%, respectively.

As seen in Table 5, such increases in the probability of guessing the correct answer yielded a relatively high mean percentage correct for an item type having fewer options. The mean percentage correct for the 3-option items

Table 4

Abbreviations for Item Types in Various Formats in Current Study

Item Types	Formats: General Test — 4- or 5-Option Multiple-Choice	Formats: Experimental Tests — 3-Option Multiple-Choice	Formats: Experimental Tests — Multiple[a]-Yes/No
Analytical Items			
Analytical Reasoning	AR5	AR3	AR2
Logical Reasoning	LR5	LR3	LR2
Numerical Logical Reasoning		NLR3	NLR2
Analysis of Explanations		AX3	AX2
Pattern Identification		PI3	
Contrasting Views		CV3	
Verbal Items			
Antonyms	ANT		
Analogies	ANAL		
Sentence Completion	SNCP		
Reading Comprehension	RCMP		
Quantitative Items			
Quantitative Comparisons	QC		
Discrete Quantitative	DQ		
Data Interpretation	DI		

[a]Multiple-yes/no is treated statistically as a 2-option item; thus, a "2" is included in its abbreviation.

Table 5

Item Characteristics for Reasoning Item Types

Type of Reasoning Item	N	Mean Percent Correct	(SD)	Reliability	Mean R-biserial	(SD)
December 1988 General Test						
AR5	38	53	(20)	.88	.54	(.11)
LR5	12	50	(20)	.63	.41	(.08)
3-Option Experimental Test						
AR3	16	73	(15)	.73	.64	(.11)
LR3	16	78	(10)	.68	.60	(.08)
NLR3	16	66	(15)	.67	.56	(.09)
AX3	16	76	(24)	.59	.60	(.14)
PI3	16	62	(22)	.84	.72	(.12)
CV3	18	68	(14)	.63	.50	(.16)
Multiple-Yes/No Experimental Test						
AR2	30	87	(8)	.77	.62	(.15)
LR2	30	80	(16)	.66	.49	(.14)
NLR2	36	79	(15)	.74	.48	(.14)
AX2	36	80	(18)	.62	.49	(.22)

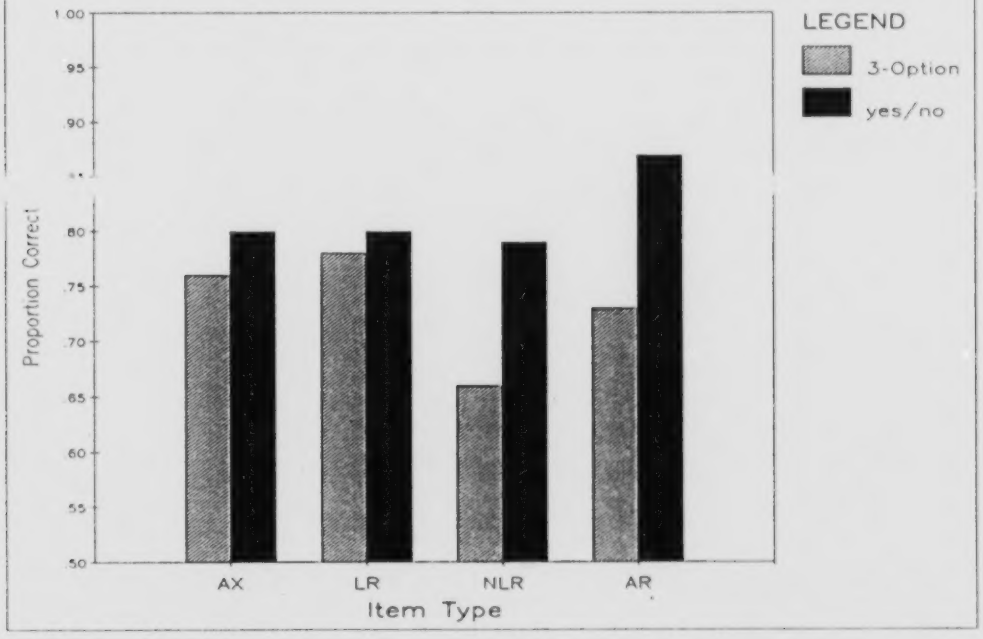
varied from 62% to 78%, while that for the multiple-yes/no items varied from 79% to 87%. Interestingly, however, decreasing the number of options did not have comparable effects on all of the item types, as seen in Figure 1. For the AR item type and for the NLR item type, the difference in mean percentage correct between the two formats was close to 17%, which is what might be expected given the difference in the probability of guessing the correct answer in each of the two formats (50% - 33% = 17%). However, for the LR item type and for the AX item type, the difference in mean percentage correct between the two formats was smaller than expected. Thus, AR and NLR appear to be about equally difficult in both formats, whereas LR and AX appear to be more difficult in a multiple-yes/no format than in a 3-option format. We cannot rule out the possibility that these observed differences arose from uncontrolled differences in how the item types happened to be translated from one format to the other. Nevertheless, findings presented in subsequent sections of this report tend to support a more substantive interpretation that will be developed in some detail.

Reliability and Discrimination. The measure of reliability presented in Table 5 is the Alpha coefficient. All other things equal, the Alpha coefficient increases in magnitude as the number of items increases or as the number of options (per item) increases. As noted earlier, we attempted to offset variations in the Alpha coefficient, as a function of the number of options, by increasing the number of items as the number of options decreased. The findings reported in Table 5 indicate that this effort was largely successful. Regardless of format, the Alpha coefficients for all the experimental item types were roughly comparable to those found for the item types currently used in the GRE analytical measure.

Included in Table 5 is the mean r-biserial correlation, a statistic that can also serve as an index of the homogeneity of a set of items. Because no prior information was available regarding how the various experimental item types would relate to one another, the criterion measure used to compute the r-biserials consisted solely of all the items of the same item type within the same battery. This procedure contrasts with that used to determine the r-biserial correlations for the analytical items from the December 1988 GRE analytical measure, where total scores on both the analytical reasoning items (AR5) and the logical reasoning items (LR5) were composited to form the criterion. Given this difference in procedure, comparisons become difficult, although it is important to note that the r-biserial correlations for the experimental item types appear to be quite acceptable (see Table 5). Also, because little is known about how the r-biserial correlation is affected by the probability of choosing the correct response by chance, or by the difficulty level of the item (see Bejar, Chaffin, & Embretson, 1991), it remains uncertain whether it is meaningful to compare r-biserial correlations across different item formats.

In the present study, the r-biserial correlations for the experimental items were used in conjunction with post-hoc item content reviews by test development staff for purposes of item quality control. The overall quality of the experimental items proved to be very high, despite the fact that most of these items had not been pretested. (All items in an operational form of the GRE General Test have passed the quality control procedures

Figure 1. Mean Proportion Correct for Item Types in Two Different Formats



accompanying pretesting). The r-biserial correlation of each experimental item was greater than zero, and those items having r-biserial correlations less than .30 were reviewed. Only one such item, a logical reasoning item in the multiple-yes/no format, was found to be ambiguous and was therefore excluded from further analyses.

Criterion-Related Validity

We explored the ability of the experimental item types to predict academic performance (criterion-related validity) by correlating the scores on the experimental item types with self-reported undergraduate grade-point averages (UGPA). Data on UGPA were obtained from the background questionnaire the examinees completed when they registered for the GRE General Test. Although self-reported UGPA has its share of shortcomings as a criterion for a graduate admissions test, we capitalized on the fact that it was readily available as an initial though tentative indicator of validity.

Table 6 presents the corrected correlations between UGPA and performance on the experimental item types and on the reasoning items from the December 1988 GRE General Test. Results for the two experimental battery samples are presented separately in the table. (The uncorrected correlations are presented in Appendix C.)

As seen in Table 6, the correlations between UGPA and the experimental item types are very similar in size to those between UGPA and the current GRE measures. The correlations do not vary extensively among the experimental item types, although the analysis of explanations (AX) item type, regardless of format, appears to be the best predictor of UGPA. These findings regarding the experimental item types are very encouraging. They suggest that the predictive validity of the analytical measure would not be reduced, and might even be improved, by incorporating some of the experimental item types into the GRE analytical measure.

Sex Differences

Also of interest were possible sex differences on the various reasoning measures. Although somewhat constrained by the relatively small sample sizes for the experimental measures, the sex comparisons did provide some useful information. Because the males in the experimental sample had higher mean scores on all three measures of the GRE General Test (Table 2), better performance by males on most if not all of the General Test item types would be expected.

Table 7 presents the effect sizes for differences between females and males in the mean proportion correct on the item types in the December 1988 GRE General Test and in the experimental batteries. As seen in Table 7, males did perform better than females on almost all of the item types, the notable exceptions being two of the experimental item types, AX2 and AX3, which appear to have negligible sex differences. Overall, the magnitudes of the sex differences on the experimental item types were comparable to those found for the reasoning item types currently in use in the GRE analytical measure. For the reasoning item types as a whole, the magnitudes of the sex differences

Table 6

Correlations of Reasoning Item Types with UGPA

Item Types	Sample of Examinees Administered Multiple-Yes/No	Sample of Examinees Administered 3-Option Multiple-Choice
December 1988 Items		
AR5	.30	.35
LR5	.35	.41
Experimental Items		
AR	.28	.28
LR	.28	.32
NLR	.36	.36
AX	.41	.38
PI		.32
CV		.34

Note. Correlations have been corrected for errors of measurement by use of the formula $r_{ab}/\sqrt{r_{aa} * r_{bb}}$. Uncorrected correlations are presented in Appendix C.

Table 7

Effect Size for Sex Differences in the Mean Proportion Correct for Item Types on the December 1988 GRE General Test and on the Experimental Test

Item Types	Sample of Examinees Administered	
	Multiple-Yes/No	3-Option Multiple-Choice
December 1988 Test		
Verbal Item Types		
ANT	-.35	-.30
ANAL	-.21	-.30
SNCP	-.35	-.19
RCMP	-.32	-.16
Quantitative Item Types		
QC	-.71	-.56
DQ	-.83	-.58
DI	-.67	-.61
Analytical Item Types		
AR5	-.32	-.11
LR5	-.40	-.35
Experimental Test		
AR	-.23	-.21
LR	-.36	-.41
NLR	-.46	-.33
AX	.00	-.07
PI		-.26
CV		-.24

Note. Effect size = $(x_f - x_m)/S_f$

were smaller than those for the quantitative item types but roughly equivalent to those for the verbal item types.

Order Effects

For the alternate forms within a given battery, the order of the blocks of items (see Table 1) was reversed. This procedure enabled us to analyze the data for order effects. The possibility of finding order effects loomed large because, in the past, otherwise acceptable reasoning item types have been found to be susceptible to practice or coaching effects (Wild, Swinton, & Wallmark, 1982). In the current study, order effects were evaluated by performing an Order (2) x Item Block (2) repeated measures analysis of variance (ANOVA) on the two blocks of a particular item type that occurred earliest and latest on the alternate test forms. For example, one block of LR3 items (Block A) appeared near the beginning of test section 1 on test Form 1 and another block (Block Z) appeared near the end of test section 3 on test Form 1 (Order 1). The positions of these two item blocks were reversed on test Form 2 (Order 2). Thus a significant interaction between Order and Item Block indicates an effect that can be characterized either as a practice effect (performance improves over time) or a fatigue-boredom effect (performance deteriorates over time).

Table 8 summarizes the magnitudes of the order effects in terms of both the order effect per item and the effect size. The order effect represents the difference in the mean proportion correct for items when they appeared in a later section of a test as opposed to when they appeared in an earlier section of a test ((Block Z, Form 1 + Block A, Form 2) - (Block Z, Form 2 + Block A, Form 1)). It will be noted that ability differences between the experimental samples that took the two different forms, as well as differences in the difficulty of the two item blocks, are counterbalanced in this analysis. A positive value indicates a practice effect while a negative value indicates a fatigue-boredom effect. The effect size is the order effect divided by the pooled standard deviation. The F value for the Order X Item Block interaction for each item type also is presented in Table 8.

Two item types--analysis of explanations in a fixed format and logical diagrams--were eliminated from an earlier version of the analytical measure, in part because of "large" practice effects of .04 to .06 per item (Wild et al., 1982). In the present study, significant Order x Item Block interactions were found for three of the item types. For two of these item types, PI3 and AX2, performance improved slightly with practice. Although the order effect per item for PI3 is of the same magnitude as those reported for the item types previously eliminated from the measure, the effect size for this item type is small by conventional standards (Cohen, 1977). Similarly, the effect size for AX2 is also small. Nevertheless, further investigation of these effects is warranted to determine whether they are persistent, whether they are altered as result of additional practice or coaching, and/or whether they would disappear as the item types become more familiar to examinees over time.

A fatigue-boredom effect, one that probably can be eliminated readily, appeared to occur in the case of NLR3. Almost all the NLR3 items were

Table 8

Order Effects for Types of Reasoning Items

Type of Reasoning Item	Sample of Examinees Administered: Multiple-Yes/No — Order Effect/ Item	Multiple-Yes/No — Effect Size	Multiple-Yes/No — Order x Item Block F(1,384)	3-Option Multiple-Choice — Order Effect/ Item	3-Option Multiple-Choice — Effect Size	3-Option Multiple-Choice — Order x Item Block F(1,372)
AR	-.01	-.04	.15	.00	.00	.00
LR	.01	.00	1.59	-.01	-.05	.85
NLR	.00	.00	.07	-.06	-.27	26.44**
AX	.01	.21	4.50*	.00	.00	.10
PI				.04	.15	11.71**
CV				.00	.00	.07

*$p < .05$

**$p < .001$

presented in the second section of the test (see Table 1). The first block of NLR3 (eight items) occurred early in the second section, and the second block was split between the end of the second section (six items) and the beginning of the third section (two items). The fact that so many items of the same kind were presented in a single section probably contributed to examinee boredom with the item type.

Relationships of the Experimental Item Types to the Current GRE Measures

We now turn to an examination of some key relationships between measures from the current GRE General Test and measures from the experimental batteries. Findings on the variance shared by different groupings of experimental item types are presented more formally in the factor analyses of Chapter 3. Here, the focus is on gaining a fuller understanding of the experimental item types, considered individually, by examining selected zero-order correlations among the measures, corrected for unreliability. (The complete correlation matrices, both uncorrected and corrected for unreliability, are found in Appendix C.)

Selected correlations are presented in Table 9, separately for the samples taking the two experimental batteries. For both samples, we see once again the familiar and problematic pattern regarding the current GRE analytical measure: the AR5 item type is more strongly related to the GRE quantitative measure than to the LR5 item type, and the LR5 item type is more strongly related to the GRE verbal measure than to AR5.

As also seen in Table 9, our experimental versions of item types currently used in the GRE analytical measure (AR3, LR3, AR2, and LR2) generally exhibited the expected patterns of correlations with the current GRE measures. For example, within the 3-option group, AR3 correlated highest with AR5, and LR3 correlated highest with LR5. Although not surprising, these and most of the other outcomes reported in Table 9 are encouraging because they indicate that most of the experimental measures would be suitable for the GRE General Test. The most noteworthy exception is AR2, which generally had the lowest correlations with the General Test measures, signaling a potential problem with the analytical reasoning item type cast in the multiple-yes/no format, an issue that is addressed further in the next chapter.

Regarding the newer experimental item types, analysis of explanations (AX3 and AX2) exhibited relationships with the current GRE General Test measures that were similar to those shown by logical reasoning (LR3, LR2). Another unsurprising outcome was that numerical logical reasoning (NLR2, NLR3) tended to correlate about equally with both the verbal and quantitative measures of the GRE General Test. On the other hand, pattern identification (PI3) presents a somewhat ambiguous picture. This item type related about equally well to AR5 and LR5, a promising sign for enhancing convergent validity, but the correlations of this item type with the General Test measures tended to be relatively low, suggesting that it may have qualities not shared with the current test. Whether such unique qualities would strengthen or weaken the GRE analytical measure cannot be determined from these data. Finally, the high correlation of contrasting views (CV3) with the

Table 9

Correlations of Types of Reasoning Items on December 1988 General Test and on Experimental Tests with Current GRE Measures

Type of Reasoning Item	3-Option Sample Current GRE Measures			
	V	Q	AR5	LR5
December 1988 Test				
AR5	.58	.82	1.00	.69
LR5	.88	.70	.69	1.00
Experimental Test				
AR3	.64	.82	.85	.72
LR3	.82	.74	.65	.92
NLR3	.75	.75	.65	.84
AX3	.77	.50	.50	.75
PI3	.50	.64	.64	.63
CV3	.93	.60	.60	.89
	Multiple-Yes/No Sample Current GRE Measures			
	V	Q	AR5	LR5
December 1988 Test				
AR5	.57	.77	1.00	.65
LR5	.88	.67	.65	1.00
Experimental Test				
AR2	.47	.65	.64	.53
LR2	.80	.66	.60	.86
NLR2	.74	.69	.64	.83
AX2	.68	.48	.58	.70

Note. Correlations have been corrected for errors of measurement by use of the formula $r_{ab}/\sqrt{r_{aa}*r_{bb}}$. Uncorrected correlations are presented in Appendix C.

GRE verbal measure suggests that this item type might be less suitable for the GRE analytical measure than for the verbal measure.

Relationships Among the Experimental Item Types

We are now in a good position to consider some possible clusterings of the experimental item types that will be among those identified more formally later in the factor analyses.

The correlations among the experimental item types are presented in Table 10. Among the 3-option multiple-choice item types, several potential groupings of item types can be discerned. Four of the item types--LR3, NLR3, AX3, and CV3--relate well with one other. LR3 and NLR3 relate especially highly with one another, they tend to relate well to all of the other item types, and they both relate about equally to the GRE verbal and quantitative measures (Table 9). As we have seen, CV3 probably would need to be eliminated from this cluster, at least as a candidate for the analytical measure, because of its high relationship with the GRE verbal measure (Table 9). Thus, we can begin to discern one possibility for the analytical measure, consisting of LR3, NLR3, and AX3. Another reasonably homogeneous cluster might consist of AR3, LR3, and NLR3. Even the combination of AR3, LR3, NLR3, and PI3 would appear to be reasonably homogeneous.

Regarding the multiple-yes/no format (Table 10), once again AR2 relates relatively poorly with the other item types, and LR2, NLR2, and AX2 would appear to form a reasonably homogeneous cluster.

Summary

Characteristics of the Experimental Item Types

The observed strengths and weaknesses of the individual experimental item types are summarized briefly below. We also offer comments on the observed effects of item format (3-option multiple choice vs. multiple-yes/no).

Analytical Reasoning. In the 3-option format (AR3), this item type functioned very much as it does in a 5-option format (AR5). AR3 correlated highly with the quantitative measure and moderately with the verbal measure. The correlation between AR3 and LR3 was higher than that between AR5 and LR5, though still not higher than that between AR3 and the quantitative measure. This pattern suggests that casting both AR and LR in a 3-option format might improve the unity of the analytical measure but probably would not enhance its discriminant validity in relation to the GRE quantitative measure. On the other hand, analytical reasoning, when cast in the multiple-yes/no format (AR2), appears to be problematic in several important respects.

Logical Reasoning. Logical reasoning worked relatively well in both the 3-option multiple choice format (LR3) and in the multiple-yes/no format (LR2). Regardless of format, logical reasoning correlated well with all the other

Table 10

Correlations Among Types of Reasoning Items on the Experimental Tests

Type of Reasoning Item	3-Option Format					
	AR3	LR3	NLR3	AX3	PI3	CV3
AR3	1.00	.80	.71	.57	.69	.58
LR3	.80	1.00	.95	.80	.64	.84
NLR3	.71	.95	1.00	.76	.70	.88
AX3	.57	.80	.76	1.00	.57	.84
PI3	.69	.64	.70	.57	1.00	.52
CV3	.58	.84	.88	.84	.52	1.00

	Multiple-Yes/No Format			
	AR2	LR2	NLR2	AX2
AR2	1.00	.72	.59	.59
LR2	.72	1.00	.84	.78
NLR2	.59	.84	1.00	.78
AX2	.59	.78	.78	1.00

Note. Correlations have been corrected for errors of measurement by use of the formula $r_{ab}/\sqrt{r_{aa}*r_{bb}}$. Uncorrected correlations are presented in Appendix C.

experimental item types. When cast in the multiple-yes/no format (LR2), this item type appeared to be relatively difficult (in relation to chance responding), suggesting that LR2 may be an especially suitable format for logical reasoning.

Numerical Logical Reasoning. This item type worked well in both the 3-option multiple-choice format (NLR3) and the multiple-yes/no format (NLR2). It was highly correlated with LR in both formats. Its correlations with the verbal and quantitative measures were comparable to each other and intermediate in magnitude. Numerical logical reasoning appears to have promise for increasing the unity of the GRE analytical measure, although it probably would not enhance the measure's discriminant validity.

Analysis of Explanations. This item type appeared to be especially successful. The relationships of both AX3 and AX2 with the verbal measure were somewhat weaker than those found for some of the other item types. Like LR, this item type was more difficult (in relation to chance responding) when cast in the multiple-yes/no format (AX2). The criterion-related validity of this item type was good in either format, and the sex differences were especially small. A small practice effect was found when this item type was cast in the multiple-yes/no format (AX2).

Pattern Identification. This item type was cast only in the 3-option multiple-choice format (PI3). Its correlations with both the verbal and quantitative measures were relatively modest, as were its correlations with the other experimental measures. The internal consistency reliability of PI3 was especially high, and this item type appeared to include some unique variance, the implications of which for the measurement of reasoning could not be determined from the correlational analyses alone. The practice effect observed for PI3 calls for further scrutiny of this item type.

Contrasting Views. This item type was cast only in the 3-option format (CV3). Because its correlation with the verbal measure was so high (.93), CV3 appears to be less well suited for the analytical measure than for a verbal measure.

Influence of Item Format on the Measurement of Reasoning Skills

One potential advantage of multiple-yes/no reasoning items is the possibility that more reasoning operations can be assessed per word (or other symbol) presented in the supporting text, thereby placing less time pressure on the examinee. Of course, such a reduction in the number of options per item would need to be accompanied by an increase in the number of items to compensate for the change in the theoretical chance parameter (say, from .20 in a 5-option format to .50 in a multiple-yes/no format). But the yes/no format would enable test developers to base as many as 10 items on a single short passage, at least in the case of AX2, perhaps resulting in significant savings in the amount of reading time required during testing. Inspection of some of the examinees' responses to our brief post-examination questionnaire provided anecdotal support for the hypothesis that testing time might be reduced by increasing the number of items while reducing the number of

options. However, it should also be noted that the present study was not designed to test this hypothesis.

Multiple-choice items and multiple-yes/no items may evoke slightly different reasoning processes. A multiple-choice item poses a comparative judgment situation in which the examinee knows that the intended correct answer is to be found among the options provided. In the case of multiple-yes/no test items, however, more than one answer can be keyed "yes," and the examinee is asked to evaluate the correctness of each statement that is presented, standing alone. Because the multiple-yes/no task opens up the possibility of having several correct answers to a question stem, or none, it permits a more natural simulation of real-life reasoning problems for which single best answers are not guaranteed to exist. The multiple-yes/no task may even pose a more challenging cognitive demand than does an otherwise comparable multiple-choice item. As we have seen, when the experimental item types were placed in a yes/no format, some appeared to be more difficult relative to multiple-choice (Figure 1).

However, our findings do not suggest that the multiple-yes/no format generally is the better way to measure reasoning skills. As noted earlier, the analytical reasoning item type cast in the multiple-yes/no format (AR2) generally correlated less strongly with related item types on the existing General Test than did any of the other experimental item types. In response to this anomaly, we reexamined the AR2 item type and concluded that a multiple-yes/no response format may not be well suited for this item type. Because the stem of an AR item often requires the application of a particular constraint or rule that leads deductively to certain outcomes, there may be more overlap and redundancy and less independence among the options. As a result, the measurement of reasoning probably is diluted, resulting in the lowered correlations.

Thus, it appears that the influence of format on the assessment of reasoning skills depends greatly on which item type is under consideration. We are suggesting that the response format interacts with the item type in such a way as to produce variation in the reasoning skill(s) (content) measured by the item type. Sometimes a particular combination of format and item type will facilitate the measurement of reasoning skills, and sometimes it will not. In this view, format does not function simply as a method-variance "main effect" that is to be isolated and eliminated from measures of reasoning. Rather, we emphasize the importance of distinguishing between those linkages of content with format that facilitate the measurement of reasoning skills and those that do not or that even weaken measurement. This issue is explored further in Chapter 3.

There is a problem that the multiple-yes/no format may pose for test development. Even though item writers attempt to formulate keys that are correct in an absolute sense, a multiple-choice format is less stringent because the formal requirement is only that the key be the best among the options provided. By contrast, a multiple-yes/no format entails a stricter criterion. Lacking a comparative context, one in which the intended correct answer can be said to be better than the alternative(s) provided, the question of whether a statement is either "correct" or "incorrect" entails a more

absolute judgment--for the item writer as well as for the examinee. At this time, then, there is some question about whether the multiple-yes/no format should be used to develop operational reasoning items.

Chapter 3

Improving the Convergent and Discriminant Validity of the GRE Analytical Measure

Rationale and Strategy

A number of the findings presented in Chapter 2 bear on the validities of the experimental item types. But we have yet to consider, more formally, whether one or more of the experimental item types would strengthen the construct validity of the GRE analytical measure. Using factor-analytic methods, we now turn to estimating how the existing measure's convergent validity and discriminant validity might be altered by substituting various combinations of the experimental item types for both analytical reasoning in the 5-option format and logical reasoning in the 5-option format.

We shall adhere to the three-dimensional design of the GRE General Test (verbal, quantitative, and analytical measures) as well as to the recommendation that this study be limited to an examination of the consequences of varying the composition of the analytical measure only (Ward et al., 1986, p. 5).

In this spirit, the 3-factor model was tested in a series of confirmatory factor analyses, differing only with respect to the particular combination of item types used to define the analytical factor. Using this procedure, supplemented by an exploratory factor analysis, we attempted to simulate what would happen to the operational analytical measure if different combinations of experimental item types were to be substituted for the existing combination (AR5 and LR5).

Even if it turned out that we could not enhance convergent validity and/or discriminant validity, the GRE General Test might still be enriched by including more facets of reasoning. From this perspective, each of the new item types was viewed as valuable in its own right--to be preserved if at all possible. However, it was also recognized that an experimental item type could fail to meet any of a number of important criteria for the GRE analytical measure, resulting in a recommendation that the item type be dropped as a contender.

Factor Analytic Procedure

The procedure for setting up the 3-factor model for each of the confirmatory factor analyses (Tables 11a, 12a, 13a, 16a, 18a, 19a) was as follows:

1. Factor loadings for the four verbal item types from the GRE General Test (ANT, ANL, SNCP, RCMP) were to be estimated on the verbal factor, but not on either of the two other factors, to which they were constrained to have zero loadings.

2. Factor loadings for the three quantitative item types from the GRE General Test (QC, DQ, DI) were to be estimated on the quantitative factor, but

not on either of the other two factors, to which they were constrained to have zero loadings.

3. Factor loadings for one or another combination of reasoning item types (AR5, AR3, AR2, LR5, LR3, LR2, NLR3, NLR2, AX3, AX2, PI3, CV3) were to be estimated on the analytical factor, but not on either of the other two factors, to which they were constrained to have zero loadings.

In applying a structural equation program (Bentler, 1985), the maximum likelihood estimation procedure (Joreskog, 1970) was used to estimate the unknown (nonzero) factor loadings from the sample covariance matrix, subject to the pattern of zero constraints and allowing the factors to be intercorrelated. In evaluating the unity (convergent validity) of the analytical measure, we report several indicators of goodness-of-fit, such as the homogeneity of its factor loadings (inspection), the Tucker-Lewis Index (e.g., Rock, Bennett, & Jirele, 1988), and the Mean Off-Diagonal Standardized Residual. In evaluating discriminant validity, we examine the resulting correlations among the three factors in the model (e.g., Stricker & Rock, 1987).

Defining Parcels

Each of the item types included in the December 1988 verbal and quantitative measures was subdivided into two parcels, yielding eight markers for the verbal factor and six markers for the quantitative factor. Because there were only two item types on the December 1988 analytical measure, they were subdivided into three parcels each rather than two in order to generate a sufficient number of markers for an analytical factor. Each of the experimental item types was divided into two parcels. Items were assigned to parcels in an alternating, sequential pattern (odd-even in the two-parcel cases). Parcels within an item type were inspected and adjusted, if necessary, to assure that the parcels were roughly equivalent in terms of mean item difficulty.

This procedure of defining the parcels within the item types represents only one of several approaches that might be used. An important consideration was that this procedure allows for the emergence of the maximum number of factors. Our intent was to explore how the measurement of reasoning skills might best be extended within the analytical measure. For this reason, an initial error on the side of overfactoring was much preferable to an underfactoring error. Moreover, an initial overfactoring error, if any, could be corrected readily. We would simply inspect the resulting factor intercorrelations defined by the item-type parcels within LISREL, and then judge whether one or more of the factors showed little or no discriminant validity.

Establishing the Base

To establish a base for the existing GRE General Test, we first ran parallel confirmatory factor analyses on our two samples, using the parcels for the General Test item types exclusively to define all three factors.

The results for the two experimental samples are reported in Table 11a. The various indicators (e.g., factor loadings, Tucker-Lewis Indexes) generally were sufficiently similar in the two samples to treat either of the two sets of outcomes as a replication of the other. This means that we will be working from essentially the same base structure in the two samples as we move on to examine various combinations of experimental item types for the analytical measure.

For both samples, the Tucker-Lewis Index indicates an adequate fit with the 3-factor model on which the GRE General Test is based (Table 11a). Also, as expected, the analytical measure is least unified: the logical reasoning parcels (LR5) have considerably lower loadings on the analytical factor than do the analytical reasoning parcels (AR5). This situation reflects the existing problem of convergent validity noted earlier. Regarding discriminant validity, the relatively high correlations between the quantitative and analytical measures probably reflect the fact that, as now constituted, the GRE analytical measure gives greater weight to the AR5 item type (76% of the items) than to the LR5 item type (24% of the items).

As a supplement to Table 11a, Table 11b presents the intercorrelations among the three measures derived by summing, for each of the three factors, the subscores for those item types that were assigned to that factor. Table 11b presents both the observed correlations and the correlations corrected for unreliability. We report these supplementary correlations because, relative to the interfactor correlations reported in Table 11a, the supplementary correlations are based on measures that are closer in their composition to the measures that would result from an actual scoring procedure. In Table 11b, the correlations corrected for unreliability are very similar to the interfactor correlations reported in Table 11a, supporting our procedure of using the latter to evaluate discriminant validity.

In reporting the results of each of the subsequent confirmatory factor analyses, we shall once again present the supplementary table of intercorrelations among the derived measures, both observed and corrected for unreliability (See Tables 12b, 13b, 16b, 18b, and 19b). However, we shall not be referring again to these supplementary tables because, in every case, the corrected correlations were very similar to the interfactor correlations from the corresponding confirmatory factor analysis (Tables 12a, 13a, 16a, 18a, and 19a, respectively).

Substitution of the 3-Option Multiple-Choice Item Types

The confirmatory factor analysis reported in Table 12a substitutes the experimental 3-option multiple-choice item types for the existing item types in the GRE analytical measure (AR5 and LR5). Compared to the base structure (Table 11a), the experimental item types are more homogeneous in their loadings on the analytical factor. In addition, the Mean Off-Diagonal Standardized Residual decreases somewhat. Both of these changes suggest potential for increasing the unity of the analytical measure, although the decrease in the Tucker-Lewis Index (compared to that seen in Table 11a) suggests that one or more of the experimental item types would not help unify

Table 11a

Confirmatory Factor Analyses: 3-Factor Solutions
GRE General Test Item Types Only

Factor Loadings for the Two Experimental Samples

Item Type and Parcel	Sample Administered the 3-Option Format			Sample Administered the Yes/No Format		
	Verbal	Quant.	Analyt.	Verbal	Quant.	Analyt.
ANT a	.72			.74		
ANT b	.81			.81		
ANL a	.67			.70		
ANL b	.65			.60		
SNCP a	.75			.74		
SNCP b	.71			.71		
RCMP a	.77			.74		
RCMP b	.70			.73		
QC a		.83			.81	
QC b		.80			.86	
DQ a		.84			.84	
DQ b		.76			.76	
DI a		.64			.65	
DI b		.62			.64	
AR5 a			.85			.83
AR5 b			.87			.86
AR5 c			.87			.84
LR5 a			.49			.56
LR5 b			.46			.43
LR5 c			.43			.38

		V	Q		V	Q
Correlations Among the Three Factors	Q	.63		Q	.54	
	A	.63	.83	A	.64	.78

For Df = 167, Chi Square = 482, $p < .001$	For Df = 167, Chi Square = 498, $p < .001$
Tucker-Lewis Index .914	Tucker-Lewis Index .911
Mean Off-Diagonal Standardized Residual .057	Mean Off-Diagonal Standardized Residual .061

Table 11b

Correlations Among the December 1988 GRE
Verbal, Quantitative, and Analytical Measures

	Sample of Examinees Administered					
	3-Option Format			Multiple-Yes/No Format		
	V	Q	A	V	Q	A
V	---	.56	.61	---	.49	.62
Q	.62	---	.75	.54	---	.71
A	.68	.84	---	.69	.80	---

Note. Values above the diagonal are observed correlations; those below are corrected for errors of measurement using the formula $r_{ab}/\sqrt{r_{aa} * r_{bb}}$.

Table 12a

Confirmatory Factor Analysis: 3-Factor Solution

	Factor Loadings		
	Measures taken from the GRE General Test		Measures taken from the 3-Option Battery
Item Type and Parcel	Verbal	Quant.	Analyt.
ANT a	.70		
ANT b	.79		
ANL a	.67		
ANL b	.66		
SNCP a	.75		
SNCP b	.71		
RCMP a	.79		
RCMP b	.72		
QC a		.83	
QC b		.81	
DQ a		.83	
DQ b		.75	
DI a		.63	
DI b		.62	
AR3 a			.65
AR3 b			.65
LR3 a			.63
LR3 b			.73
NLR3 a			.69
NLR3 b			.67
PI3 a			.65
PI3 b			.63
AX3 a			.47
AX3 b			.60
CV3 a			.58
CV3 b			.64

Correlations Among the Three Factors	V	Q
Q	.63	
A	.84	.80

For Df = 296, Chi Square = 850, $p < .001$
Tucker-Lewis Index: .881
Mean Off-Diagonal Standardized Residual: .048

Note. The verbal and quantitative item-type measures are from the GRE General Test, and the analytical item-type measures are from the 3-option multiple-choice experimental battery.

Table 12b

Correlations Between the December 1988 GRE Verbal and Quantitative Measures and a Simulated Analytical Measure: AR3 + LR3 + NLR3 + PI3 + AX3 + CV3

	V	Q	A
V	---	.56	.75
Q	.62	---	.71
A	.82	.79	---

Note. Values above the diagonal are observed correlations; those below are corrected for errors of measurement using the formula $r_{ab}/\sqrt{r_{aa}*r_{bb}}$.

the GRE analytical measure. (We return later to the question of which of the experimental item types could maximize unity.) Regarding discriminant validity, it will be noted that the correlation between the verbal factor and the experimental analytical measure (Table 12a) increased relative to that in the existing GRE General Test (Table 11a).

Substitution of the Multiple-Yes/No Item Types

The confirmatory factor analysis reported in Table 13a substitutes all the experimental multiple-yes/no item types for the existing item types in the GRE analytical measure. Compared to the base structure (Table 11a), the new item types are more homogeneous in their loadings on the analytical factor, the Tucker-Lewis Index increases somewhat, and the Mean Off-Diagonal Standardized Residual decreases. Here we see considerable promise for increased unity of the GRE analytical measure. Once again, the correlation between the verbal and experimental analytical measure (Table 12a) increased relative to that in the existing GRE General Test (Table 11a).

Multitrait-Multimethod Analyses

To further explore the concept of item type x format interaction discussed earlier, we subjected selected data to a multitrait-multimethod analysis. (The notion of "trait," embedded in the term "multitrait," simply refers here to an item type.) The design of the present study did not allow us to cross three traits with three methods, the minimum for uniquely identifying a multitrait-multimethod (MM) model (e.g., Bentler & Lee, 1979; Werts, Linn, & Joreskog, 1972). "Identification" can occur when there are sufficient independent equations to estimate uniquely each of the unknown parameters, and an "overidentified" system of equations is desirable because that would allow additional degrees of freedom for testing the model's fit.

Models having either too few methods or too few traits can sometimes be identified by fixing some of the unknown parameters based on information outside of the model. Another approach is to carry out preliminary tests of selected assumptions related to MM model parameters and, if they hold, to incorporate them into the model, thereby reducing the number of unknown parameters and making the model identified or overidentified (e.g., Bramble & Wiley, 1974). We used the latter approach. Because the present parcels were within item types and were constructed to be reasonably parallel, we could expect pairs of parcels sharing the same method and trait to be tau-equivalent, that is, to have equivalent factor loadings, though not necessarily equal errors of measurement. (Such an assumption is weaker than that of parallelism.)

For the sample administered, the experimental 3-option multiple-choice format, a 4-factor solution was defined by crossing the two formats with analytical reasoning and logical reasoning (AR5, AR3, LR5, LR3). These preliminary results indicated that two of the three AR5 parcels had almost identical loadings. Similarly, two of the three LR5 parcels had virtually identical loadings. Also, the two AR3 parcels had almost identical loadings,

Table 13a

Confirmatory Factor Analysis: 3-Factor Solution

	Factor Loadings		
	Measures Taken from the GRE General Test		Measures taken from the Yes/No Battery
Item Type and Parcel	Verbal	Quant.	Analyt.
ANT a	.74		
ANT b	.81		
ANL a	.70		
ANL b	.60		
SNCP a	.75		
SNCP b	.71		
RCMP a	.74		
RCMP b	.73		
QC a		.81	
QC b		.87	
DQ a		.84	
DQ b		.75	
DI a		.65	
DI b		.65	
AR2 a			.59
AR2 b			.61
LR2 a			.65
LR2 b			.70
NLR2 a			.70
NLR2 b			.72
AX2 a			.58
AX2 b			.57

Correlations Among the Three Factors	V	Q
Q	.54	
A	.78	.73

For Df = 206, Chi Square = 517, $p < .001$
Tucker-Lewis Index: .919
Mean Off-Diagonal Standardized Residual: .043

Note. The verbal and quantitative item-type measures are from the GRE General Test, and the analytical item-type measures are from the two-option yes/no experimental battery.

Table 13b

Correlations Among the December 1988 GRE Verbal and Quantitative Measures and a Simulated Analytical Measure: AR2 + LR2 + NLR2 + AX2

	V	Q	A
V	---	.49	.68
Q	.54	---	.63
A	.76	.71	---

Note. Values above the diagonal are observed correlations; those below are corrected for errors of measurement using the formula $r_{ab}/\sqrt{r_{aa}*r_{bb}}$.

and the two LR3 parcels were sufficiently close in size to assume that they are indeed tau-equivalent measures. In summary, it appears that there are two tau-equivalent measures to "mark" each of the four factors defined by crossing the two formats (methods) with the two item types (traits).

Given the above information about the tau-equivalence of the comparable markers, we can weakly overidentify the model by constraining the measures within pairs to have equal factor loadings (tau-equivalence) and by assuming independence of the method and trait factors. This procedure was carried out in both the 3-option multiple-choice sample and the multiple-yes/no sample. The results obtained from fitting the MM models are presented in Table 14 (3-option format) and Table 15 (multiple-yes/no format).

As seen in Table 14, the method variance for the 3-option multiple-choice parcels is relatively small, though significant. For example, only 14% of the common variance of LR3a is method variance $((.24^2)/(.60^2 + .24^2))$. For AR3a, about 29% of the common variance is method variance. In Table 15, the method variance for the AR2 parcels is substantial: the common variance appears to be shared about equally by trait and method. On the other hand, relatively little method variance (10%) is attributable to the LR2 parcels. Overall, it appears that the method variance associated with item format is considerably more important for the analytical reasoning item type than for the logical reasoning item type. This difference is particularly marked for the multiple-yes/no format.

Implications for the Multiple-Yes/No Item Types

The above method-variance analyses, together with findings presented earlier, lead us to suggest that AR2 be dropped as a contender for the GRE analytical measure.

These analyses also suggest, once again, that response format interacts with item type to produce variation in the reasoning skill(s) measured by the item type. Specifically, these analyses suggest that, in contrast to the analytical reasoning item type, the logical reasoning item type measures comparable (though not identical) aspects of reasoning when cast in the LR5, LR3, and LR2 formats. Of course, the method-variance analyses did not include numerical logical reasoning and analysis of explanations because these item types are not now part of the GRE General Test. However, we can test whether an analytical measure consisting of LR2, NLR2, and AX2, but not AR2, is likely to be more unified than what we could expect for an analytical measure that includes AR2, the latter already reported in Table 13a.

Table 16a provides the results of such a test. As seen in Table 16a, the combination of LR2, NLR2, and AX2 yields an excellent fit with the desired three-factor model: the Tucker-Lewis Index almost reaches .95, and the Mean Off-Diagonal Standardized Residual drops to .04. Regarding discriminant validity, comparison of Tables 11a and 16a suggests the following tradeoff: the possible analytical measure under consideration here probably would have a higher correlation with the GRE verbal measure, but also a lower correlation with the GRE quantitative measure.

Table 14

Analysis of a 2-Trait by 2-Method Model for the 3-Option Multiple-Choice Experimental Sample

	Factor Loadings			
	Trait		Method	
Item Type and Parcel	AR	LR	5-Option	3-Option
AR5a	.82		.24	
AR5b	.83		.25	
LR5a		.62	-.22	
LR5b		.66	-.24	
AR3a	.65			.41
AR3b	.67			.42
LR3a		.60		.24
LR3b		.68		.27

Correlation between Trait AR and Trait LR = .77

For Df = 19, Chi-Square = 16, $p < .63$

Tucker-Lewis Index: 1.00

Mean Off-Diagonal Standardized Residual: .02

Table 15

Analysis of a 2-Trait by 2-Method Model for the Multiple-Yes/No Experimental Sample

	Factor Loadings			
	Trait		Method	
Item Type and Parcel	AR	LR	5-Option	Multiple-Yes/No
AR5a	.72		.35	
AR5b	.82		.40	
LR5a		.51	-.03	
LR5b		.55	-.03	
AR2a	.57			.58
AR2b	.55			.56
LR2a		.63		.21
LR2b		.69		.23

Correlation between Trait AR and Trait LR = .74

For Df = 19, Chi-square = 30, $p < .05$

Tucker-Lewis = .98

Mean Off Diagonal Standardized Residual: .04

Table 16a

Confirmatory Factor Analysis: 3-Factor Solution

	Factor Loadings		
	Measures Taken from the GRE General Test		Measures Taken from the Yes/No Battery
Item Type and Parcel	Verbal	Quant.	Analyt.
ANT a	.74		
ANT b	.81		
ANL a	.70		
ANL b	.60		
SNCP a	.75		
SNCP b	.71		
RCMP a	.74		
RCMP b	.73		
QC a		.81	
QC b		.87	
DQ a		.84	
DQ b		.75	
DI a		.65	
DI b		.65	
AR2 a			Excluded
AR2 b			Excluded
LR2 a			.64
LR2 b			.70
NLR2 a			.72
NLR2 b			.74
AX2 a			.59
AX2 b			.57

		V	Q
Correlations Among the Three Factors	Q	.54	
	A	.81	.69

For Df = 167, Chi Square = 342, $p < .001$
Tucker-Lewis Index: .949
Mean Off-Diagonal Standardized Residual: .038

Note. The verbal and quantitative item-type measures are from the GRE General Test, and the analytical item-type measures, excluding analytical reasoning (AR-2), are from the two-option yes/no experimental battery.

Table 16b

Correlations Among the December 1988 GRE Verbal and Quantitative Measures and a Simulated Analytical Measure: AR2 + LR2 + NLR2 + AX2

	V	Q	A
V	---	.49	.69
Q	.54	---	.58
A	.79	.67	---

Note. Values above the diagonal are observed correlations; those below are corrected for errors of measurement using the formula $r_{ab}/\sqrt{r_{aa}*r_{bb}}$.

Further Examination of the Structure of the Reasoning Domain

We now return to the question of how experimental item types might be selected from the 3-option multiple-choice battery so as to maximize the unity of the GRE analytical measure. To shed further light on this matter, we conducted an exploratory factor analysis on all the item types administered to the 3-option sample. Using Promax, we rotated the four factors (principal components) for which the eigenvalues were greater than 1.00.

Table 17a presents the resulting factor loadings, and Table 17b presents the interfactor correlations as well as the variance explained by each of the four factors. The reader is reminded that, for a variety of reasons, the factor labels, factor loadings, and interfactor correlations reported here are not directly comparable to those contained in any of the other tables in this report, such as the confirmatory factor analyses and the correlation matrices.

As seen in Table 17a, two of the resulting factors can be identified as verbal and quantitative factors, although, as will be clarified shortly, these two factors are not equivalent to the GRE General Test measures having the same labels.

The remaining two factors divide the various analytical item types into two major subcategories. We shall call these two subcategories informal reasoning and formal-deductive reasoning. This distinction between two dimensions of reasoning parallels, to some extent, distinctions currently being drawn in cognitive psychology, in philosophy, and in education. These distinctions include such contrasts as those drawn between well-structured and ill-structured problems (Frederiksen, 1983; Simon, 1978; Ward et al., 1983), informal and formal reasoning (Scriven, 1976; Tucker, 1985; Voss, Perkins, & Segal, in press), critical thinking and formal logic (Ennis, 1987), and everyday and formal reasoning (Galotti, 1989).

Of special interest is the possibility that the four rotated factors have a particular order: verbal, informal reasoning, formal-deductive reasoning, and quantitative. The grounds for suggesting this particular order (or its reverse) are as follows:

a. It is well established that verbal and quantitative abilities tend to be moderately but not highly correlated with each other, making it reasonable to place these two factors at the two ends of this continuum.

b. Informal reasoning typically deals with the manipulation of verbal symbols as meanings embedded within a semantic network, whereas formal-deductive reasoning deals primarily with the logical manipulation of symbols as counters, often in numerical form (Tucker, 1985). It is, therefore, reasonable to suppose that informal reasoning skills are more closely linked to verbal ability than to quantitative ability, whereas formal-deductive skills are more closely linked to quantitative ability than to verbal ability.

Table 17a

Exploratory Factor Analysis of All Item Types and Parcels
For the 3-Option Multiple-Choice Experimental Sample*

General Test	Factor Loadings** Verbal	Informal Reasoning	Formal-Deductive Reasoning	Quant.
ANT a	.94	-.31	.08	-.03
ANT b	.88	-.21	.09	.06
ANL a	.74	-.04	.02	-.01
ANL b	.72	-.01	-.07	.05
SNCP a	.71	.06	-.07	.11
SNCP b	.62	.08	.01	.10
RCMP a	.54	.24	.16	-.04
RCMP b	.55	.16	.32	-.22
QC a	.07	.04	.46	.38
QC b	-.03	.07	.44	.41
DQ a	.06	-.06	.40	.53
DQ b	-.14	.04	.39	.53
DI a	.04	-.05	-.01	.82
DI b	.02	-.06	.03	.79
AR5 a	.02	-.04	.87	.00
AR5 b	.05	-.08	.90	-.02
AR5 c	.00	.02	.79	.07
LR5 a	.27	.03	.31	.08
LR5 b	.38	.30	-.10	.14
LR5 c	.46	.27	.06	-.12
Experimental Battery				
AR3 a	.06	.09	.71	-.06
AR3 b	.06	.05	.70	.00
LR3 a	.29	.37	.08	.02
LR3 b	.27	.39	.02	.22
NLR3 a	.11	.53	-.09	.32
NLR3 b	.11	.62	-.21	.29
PI3 a	-.24	.75	.33	-.07
PI3 b	-.29	.76	.32	-.03
AX3 a	.24	.63	-.19	-.13
AX3 b	.36	.51	.09	-.28
CV3 a	.57	.18	-.15	.12
CV3 b	.53	.29	.02	-.08

*Principal Components with Promax Rotation.
**Loadings equal to or greater than .30 are underlined.

Table 17b

Exploratory Factor Analysis of All Item Types and Parcels for the 3-Option Multiple-Choice Experimental Sample

	Interfactor Correlations			
	Verbal	Informal Reasoning	Formal-Deductive Reasoning	Quant.
Verbal	1.00	.59	.51	.47
Informal Reasoning	.59	1.00	.60	.51
Formal-Deductive Reasoning	.51	.60	1.00	.62
Quant.	.47	.51	.62	1.00
Factor Variance Explained	6.51	3.39	5.45	3.12

c. The patterning of the interfactor correlations (Table 17b) resembles a simplex, providing empirical support for the suggested ordering of the four factors.

As already noted, the verbal and quantitative factors identified in Table 17a are not to be construed as equivalent to the similarly labeled measures of the GRE General Test. In contrast to the method of deriving the verbal and quantitative scores on the General Test, the factor loadings that defined the factors in Table 17a weighted the item types differentially, and no a priori constraints were placed on the item types that were allowed to define the factors. Indeed, the absence of such constraints is one of the reasons why Tables 17a and 17b seem to us to be especially compelling for purposes of inferring the underlying structure of the reasoning domain.

For example, consider the factor loadings on the verbal factor reported in Table 17a, especially the loadings for the four verbal item types from the GRE General Test, having the highest loadings on this factor. The magnitudes of the loadings for these four item types had a particular order, being highest for ANT, next highest for ANL, next highest for SNCP, and lowest for RCMP. This ordering of the loadings gives especially heavy weight to the lexical or word-knowledge components of the GRE verbal measure (ANT and ANL), and does so at some expense to the comprehension and inferential components of the GRE verbal measure (SNCP and RCMP). At the same time, the ordering of the magnitudes of the loadings for the same four item types on the informal reasoning factor is precisely the reverse of that for the verbal factor! This pattern of outcomes supports the implication that the verbal and informal reasoning factors uncovered by our analysis represent closely related but distinctive domains.

With regard to the quantitative factor, a parallel situation emerges that is no less interesting (Table 17a). Here, the rank ordering of the loadings of the GRE quantitative item types is, from highest, to lowest: DI, DQ, and QC. This particular ordering begins to make sense if the quantitative factor identified by the exploratory factor analysis is defined in terms of knowledge of mathematical symbols and notations. To complete the parallel, it will be noted once again that the rank ordering of the magnitudes of the DI, DQ, and QC loadings on the formal-deductive reasoning factor is precisely the reverse of that noted for the quantitative factor. Mathematicians probably would concur that the DI and DQ item types measure formal-deductive reasoning skills as well as knowledge of mathematical symbols.

In sum, we suggest that the 4-factor structure of Tables 17a and 17b provides a compelling representation of two underlying kinds of reasoning skills together with their closely linked but distinctive knowledge or symbol systems.

Implications for Unifying the GRE Analytical Measure

The above findings on the structure of the reasoning domain have important ramifications for the analytical measure of the GRE General Test. For one thing, the implication that the reasoning domain is divided into two subdomains helps explain why it has been so difficult to construct a single

unified analytical measure for the GRE General Test. (See also Rock, Bennett, & Jirele, 1988; and Schaeffer & Kingston, 1988).

Indeed, our findings suggest that a single analytical measure for the GRE General Test will be unified to the extent that the measure includes either informal reasoning tasks or formal-deductive reasoning tasks, but not both kinds of tasks. Specifically, inspection of the factor loadings for informal reasoning and for formal-deductive reasoning (Table 17a) reveals remarkably little item-type overlap between these two factors. The one possible exception is pattern identification (PI3), the only item type in Table 17a that consistently had at least moderate loadings on both of the reasoning factors.

Status of Pattern Identification (PI3)

Of the various item types included in this study, pattern identification was least expected to exhibit properties that would make it a strong contender for a revised analytical measure. Nevertheless, as just noted, pattern identification apparently has the desirable property of providing a bridge between the two reasoning factors (Table 17a). More surprising, perhaps, was the fact that males did not outperform females on pattern identification to the extent that might have been expected (Table 7). From these standpoints, then, pattern identification appears to be a promising item type for the GRE analytical measure.

Nevertheless, we need to proceed cautiously in evaluating pattern identification as a possible contender for the GRE analytical measure. Some major considerations are as follows:

a. The study's experimental procedure provided a 6-minute period for examinees to master the instructions for this item type. From the standpoint of measuring reasoning, a case can be made for incorporating such a learning period into the GRE General Test itself. Whether this step would be feasible from an operational standpoint remains to be determined, however.

b. As reported earlier, there was evidence for a learning effect for this item type. This learning effect might be overcome by giving the examinee one or two sample pattern identification items to solve before the examinee proceeds to the actual test items. Or, the learning effect might dissipate over the years as more examinees become familiar with the relevant descriptive material in the GRE Information Bulletin. At the moment, however, we do not know whether either of these approaches would reduce the learning effect.

c. We have seen that pattern identification shared considerable variance with the other experimental item types (Table 17a). Nevertheless, the between-parcel standardized residual for pattern identification was especially high (.37) in the confirmatory factor analysis (for which other findings are reported in Table 12a), indicating that part of the reliable variance for this item type was unique (i.e., cannot be attributed to the verbal, quantitative, or analytical factor). The nature of this additional source of reliable unique variance remains unknown. Perhaps it is the result of individual differences in the degree or rate of mastery of the instructions during the

learning period. If so, the excessive residual noted for this item type would be expected to disappear if the learning effect noted in (b) above were to become neutralized.

In summary, the pattern identification item type appears to bridge the two reasoning factors, an important feature, but the status of this item type remains tentative pending further analyses of its properties and feasibility for the GRE General Test.

Status of Analysis of Explanations (AX3 and AX2)

For the various reasons already noted, analysis of explanations appears to be an especially promising contender for the GRE analytical measure. For example, this item type had the highest correlation with UGPA (Table 6), and performance on this item type was least associated with sex (Table 7). However, further consideration needs to be given to the following:

a. The study's experimental procedure provided a 6-minute period for examinees to master the instructions for this item type and to practice on a set of items. As already suggested, a case can be made for incorporating such a learning period into the GRE General Test itself.

b. The internal consistency reliabilities for analysis of explanations were satisfactory, although they were the lowest among the experimental item types (Table 5). Also, this item type typically had acceptable but relatively low factor loadings on the analytical factor in the 3-factor models (e.g., Tables 12a, 13a, 16a). Of course, increasing the number of questions asked (test items) can be expected to raise the reliabilities and perhaps also the factor loadings.

c. As reported earlier, there was a small but statistically significant learning effect for this item type. It seems likely, though not certain, that such a learning effect would dissipate over the years as more examinees become familiar with the relevant descriptive material in the GRE Information Bulletin.

Earlier we noted why the multiple-yes/no response format appears to be especially well suited to analysis of explanations. In order to capitalize on this apparently favorable linkage, we suggest that, for analysis of explanations, AX2 be given special consideration. On the other hand, analysis of explanations in the 3-option multiple-choice format (AX3) also remains a viable contender for the GRE analytical measure.

Status of Contrasting Views (CV3)

Inspection of the factor loadings for contrasting views in the exploratory factor analysis (Table 17a) supports our earlier suggestion that this item type may be a more promising candidate for a verbal measure than for the GRE analytical measure. Indeed, because contrasting views appears to be so appropriate for measuring humanities-related thinking skills, this item type might be worth considering as a candidate for the GRE verbal measure.

Maximizing the Unity of the GRE Analytical Measure

The findings of this study suggest an important conclusion. If our goals were to measure informal reasoning and formal-deductive reasoning and, in addition, to maximize the convergent validity of the GRE General Test, we would need to move to a solution in which the General Test has four (rather than three) measures: verbal, informal reasoning, formal-deductive reasoning, and quantitative.

On the other hand, the confirmatory factor analyses also support the conclusion that the unity of the 3-factor model could be improved by means of any of a number of combinations of item types. We now turn to a consideration of some of these possible combinations. We also wish to emphasize that there are important additional considerations, not included here, that would need to be addressed and resolved in discussing the future composition of the GRE analytical measure.

A-1. A radical approach would be to include only those item types having at least moderate loadings on formal-deductive reasoning (Table 17a). Such a reconstituted analytical measure might consist exclusively of analytical reasoning (AR5 and/or AR3). However, would there be much appeal in a reasoning measure calling for even less informal reasoning than does the currently operational GRE analytical measure?

A-2. A possible variant of this approach would be to include pattern identification (PI3) as well as analytical reasoning, because pattern identification appears to include both formal-deductive reasoning and informal reasoning (Table 17a). However, as suggested earlier, the viability of this variant might depend on additional information regarding the status of the pattern identification item type.

B-1. Another way to unify the analytical measure would be to include only those item types having at least moderate loadings on the informal reasoning factor (Table 17a). This approach would extend the GRE analytical measure in the sense that it would include three item types rather than two, and almost certainly this approach would enhance the measure's unity. We already have seen that, by using LR2, NLR2, and AX2, the analytical measure probably would become well unified (Table 16a). Similarly, as seen in Table 18a, the analytical measure probably would be well unified by including only LR3, NLR3, and AX3. Also, because the results reported in Tables 16a and 18a are so similar, any combination of the two formats for these particular item types appears likely to unify the GRE analytical measure.

For these reasons, B-1 becomes a very attractive alternative. There is a drawback, however. The analytical reasoning item type would then be dropped altogether from the GRE analytical measure. Some people may be concerned about the validity of an analytical measure that calls for so little formal-deductive reasoning. Also, because the currently operational GRE analytical measure relies so heavily on analytical reasoning (AR5), dropping this item type altogether from the GRE analytical measure might be seen as unacceptably abrupt.

Table 18a

Confirmatory Factor Analysis: 3-Factor Solution

Item Type and Parcel	Factor Loadings: Measures Taken from the GRE General Test Verbal	Quant.	Measures taken from the 3-Option Battery Analyt.
ANT a	.71		
ANT b	.80		
ANL a	.67		
ANL b	.66		
SNCP a	.75		
SNCP b	.71		
RCMP a	.79		
RCMP b	.71		
QC a		.83	
QC b		.81	
DQ a		.84	
DQ b		.75	
DI a		.64	
DI b		.63	
AR3 a			Excluded
AR3 b			Excluded
LR3 a			.65
LR3 b			.75
NLR3 a			.71
NLR3 b			.68
PI3 a			Excluded
PI3 b			Excluded
AX3 a			.48
AX3 b			.63
CV3 a			Excluded
CV3 b			Excluded

Correlations Among the Three Factors	V	Q
Q	.63	
A	.84	.74

For Df = 167, Chi Square = 352, $p < .001$
Tucker-Lewis Index: .944
Mean Off-Diagonal Standardized Residual: .038

Note. The verbal and quantitative item-type measures are from the GRE General Test, and the analytical item-type measures, excluding analytical reasoning (AR-3), pattern identification (PI-3), and contrasting views (CV-3), are from the three-option multiple-choice experimental battery.

Table 18b

Correlations Among the December 1988 GRE Verbal and Quantitative Measures and a Simulated Analytical Measure: LR3 + NLR3 + AX3

	V	Q	A
V	---	.56	.72
Q	.62	---	.63
A	.82	.72	---

Note. Values above the diagonal are observed correlations; those below are corrected for errors of measurement using the formula $r_{ab}/\sqrt{r_{aa}*r_{bb}}$.

B-2. A possible variant of this approach would be to drop analytical reasoning from the GRE analytical measure, but to add this item type to the GRE quantitative measure. Such a move would be supported by the factor structure reported in Table 17a. Adding analytical reasoning to the quantitative measure would have additional advantages: (a) the analytical reasoning item type would thereby be retained as part of the General Test; (b) the quantitative measure might become more accessible to examinees whose undergraduate majors are in fields other than mathematics, the physical sciences, or related areas; and (c) the considerable sex differences (favoring males) now exhibited by the quantitative measure probably would be reduced somewhat by including analytical reasoning in the quantitative measure.

Combining the Two Kinds of Reasoning

The above combinations of item types would likely maximize the unity of a reconstituted GRE analytical measure. By contrast, a strategy that attempts to include both informal reasoning and formal-deductive reasoning within a single measure will sacrifice some of that unity. Yet our findings also suggest that some combinations of item types that "cross" the two kinds of reasoning, such as the following, might provide somewhat greater unity than does the existing GRE analytical measure.

C-1. This combination would include the multiple-yes/no format applied to analysis of explanations (AX2), logical reasoning (LR2), and numerical logical reasoning (NLR2), and the 3-option multiple-choice format applied to analytical reasoning (AR3) and pattern identification (PI3).

C-2. This combination would be the same as C-1 except that pattern identification (PI3) would be excluded.

D-1. This combination would include the multiple-yes/no format applied to analysis of explanations (AX2) only, and the 3-option multiple-choice format applied to analytical reasoning (AR3), logical reasoning (LR-3), numerical logical reasoning (NLR3), and pattern identification (PI3).

D-2. This combination would be the same as D-1 except that pattern identification (PI3) would be excluded.

E-1. This combination would include the 3-option multiple-choice format only, applied to analytical reasoning (AR3), logical reasoning (LR3), numerical logical reasoning (NLR3), analysis of explanations (AX3), and pattern identification (PI3).

E-2. This combination would be the same as E-1 except that pattern identification (PI3) would be excluded.

To provide an example of the degree of unity provided by "crossing" item types, we ran a confirmatory factor analysis in which the analytical factor was defined by four of the five item types included in combination D-1 above.

As seen in Table 19a, compared to the existing GRE analytical measure (Table 11a), the experimental item types are more homogeneous in their

loadings and the Mean Off-Diagonal Standardized Residual is considerably lower. On the other hand, the Tucker-Lewis Index is lower, though only slightly. Of course, because this particular solution crosses informal reasoning (LR3, NLR3, and PI3) with formal-deductive reasoning (AR3, PI3), it cannot be expected to attain as much unity as one of the purer combinations, such as B-1 (see Tables 16a and 18a). However, if combining the two types of reasoning within a single measure were to be given higher priority than factorial purity, it appears likely that any of the solutions suggested above, and perhaps others, would yield an acceptable degree of unity for the GRE analytical measure.

Table 19a

Confirmatory Factor Analysis: 3-Factor Solution

	Factor Loadings		
Item Type and Parcel	Measures Taken from the GRE General Test Verbal	Quant.	Measures taken from the 3-Option Battery Analyt.
ANT a	.71		
ANT b	.80		
ANL a	.67		
ANL b	.66		
SNCP a	.75		
SNCP b	.71		
RCMP a	.79		
RCMP b	.71		
QC a		.83	
QC b		.81	
DQ a		.83	
DQ b		.75	
DI a		.63	
DI b		.63	
AR3 a			.68
AR3 b			.69
LR3 a			.62
LR3 b			.72
NLR3 a			.68
NLR3 b			.65
PI3 a			.69
PI3 b			.67
AX3 a			Excluded
AX3 b			Excluded
CV3 a			Excluded
CV3 b			Excluded

Correlations Among the Three Factors	V	Q
Q	.63	
A	.76	.84

For Df = 206, Chi Square = 608, $p < .001$
Tucker-Lewis Index: .899
Mean Off-Diagonal Standardized Residual: .043

Note. The verbal and quantitative item-type measures are from the GRE General Test, and the analytical item-type measures, excluding analysis of explanations (AX-3) and contrasting views (CV-3), are from the 3-option multiple-choice experimental battery.

Table 19b

Correlations Among the December 1988 GRE Verbal and Quantitative Measures and a Simulated Analytical Measure: AR3 + LR3 + NLR3 + PI3

	V	Q	A
V	---	.56	.67
Q	.62	---	.74
A	.74	.82	---

Note. Values above the diagonal are observed correlations; those below are corrected for errors of measurement using the formula $r_{ab}/\sqrt{r_{aa}*r_{bb}}$.

Chapter 4

Summary and Conclusions

The study identified, developed, and evaluated a group of experimental item types that were designed to broaden the current analytical measure of the GRE General Test, and to strengthen its construct validity. A number of criteria guided the choice of the experimental item types. Special attention was given to the measurement of a variety of aspects of reasoning, but without introducing essentially new symbolic materials, such as complex figural stimuli or options. After considering these criteria in relation to a variety of possible item types, six item types were selected for further investigation.

The selected item types were analytical reasoning and logical reasoning, the two item types currently in use in the GRE analytical measure; numerical logical reasoning, a variant of the logical reasoning item type in which the stimulus includes a table, a graph, or other quantitative material; analysis of explanations (revised), a variant of an item type formerly included in the GRE analytical measure; pattern identification, a constrained version of the number series item type found in the psychological literature; and contrasting views, believed to be especially suitable for measuring aspects of reasoning applied to content from the humanities.

For both substantive and efficiency reasons, there was also interest in exploring possible alternatives to the exclusive use of the 5-option multiple choice format in the GRE analytical measure. It was thought that a multiple-yes/no format might place somewhat different cognitive demands on examinees as they attempt to answer reasoning questions, thereby contributing to a broadening of the current GRE analytical measure. From an efficiency standpoint, it was thought that using either a 3-option multiple-choice format or a multiple-yes/no format might reduce time pressures on examinees without reducing efficiency of measurement, although this hypothesis was not actually tested in the study.

As a result of these considerations, all six of the experimental item types were developed in a 3-option multiple-choice format, and four of them also were developed in a multiple-yes/no format. Two experimental batteries were assembled, one using the 3-option format and the other using the multiple-yes/no format. In order to investigate practice and fatigue-boredom effects, two forms of each of the experimental batteries were developed, one form presenting the same items as the other but in approximately the reverse sequence. Two samples of approximately 370 examinees each, all of whom had recently taken the GRE General Test, were given or the other experimental batteries.

Each of the experimental item types, whether cast in a multiple-yes/no or a 3-option multiple-choice format, had at least some promising features. Each appeared to have the capacity to contribute satisfactorily to internal consistency reliability. Sex differences on the experimental item types generally were within the same range as sex differences on the item types currently in use, and analysis of explanations exhibited especially small sex

differences. The correlations of the experimental item types with self-reported undergraduate grade-point average generally were similar and occasionally slightly higher than were the comparable correlations for the current measures of the GRE General Test.

A fatigue-boredom effect in the case of numerical logical reasoning appeared to be an artifact of the way that this item type happened to be placed within the experimental battery. A small practice effect was observed for pattern identification, although the possibility remains that this effect could be largely eliminated by providing greater opportunities for examinees to work on practice items during the test session. Contrasting views appeared to be more suitable as part of a verbal measure than as part of the GRE analytical measure.

Regarding item format, the 3-option multiple-choice format appeared to be satisfactory for all of the experimental item types. Several lines of evidence suggested that substituting the multiple-yes/no format for a multiple-choice format might result in the assessment of different aspects of reasoning, although the evidence also suggested that such an effect would be differentially applicable to the experimental item types. For example, the multiple-yes/no format appeared to be especially suitable for analysis of explanations (revised), but this particular format proved to be unsuitable for analytical reasoning.

Different combinations of the experimental item types were evaluated in a series of confirmatory factor analyses that simulated what would happen if a particular combination were substituted for the item types used currently in the GRE analytical measure. There was evidence that certain combinations of the experimental item types would broaden the GRE analytical measure and strengthen its convergent validity. On the other hand, the results suggested that most if not all of the combinations would be unlikely to improve the discriminant validity of the GRE analytical measure.

The reported analyses of different possible combinations of experimental item types could serve as partial guidelines for reconsidering the composition of the GRE analytical measure. However, the present study was not designed to answer some questions that would need to be resolved before any of the experimental item types are incorporated into the analytical measure of the GRE General Test.

The study's findings provided evidence for the view that the reasoning domain consists of two major subdomains: informal reasoning and formal-deductive reasoning. This outcome helped explain why it has proven to be so difficult to unify the GRE analytical measure. Yet it also provided, for the first time, quite precise theoretical guidelines for selecting those combinations of experimental item types that would be most likely to maximize the unity of the GRE analytical measure. The hypothesis that informal reasoning and formal-deductive reasoning constitute related but distinct reasoning subdomains appears to be especially worthy of further conceptual elaboration and empirical investigation.

The present study opens up new possibilities for strengthening the analytical measure of the GRE General Test. However, in accordance with an earlier recommendation (Ward et al., 1986), the present report was limited to possible near-term changes in the GRE analytical measure only. Of course, the findings reported here, together with additional simulations based on the present and other data, could be used to consider alternative combinations of item types for the verbal, quantitative, and analytical measures of the GRE General Test. Addressing the possibility of strengthening the GRE General Test as a whole would be a natural extension of both the spirit and substance of the present study.

References

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. Journal of Educational Measurement, 25, 111-123.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991) Cognitive and psychometric analysis of analogical problem solving. New York: Springer-Verlag.

Bentler, P.M. (1985). Theory and implementation of EQS and EQS/PC. Unpublished manuscript. Los Angeles: BMDP Statistical Software.

Bentler, P. M., & Lee, S. Y. (1979). A statistical development of three mode factor analysis. British Journal of Mathematical and Statistical Psychology, 32, 87-104.

Bramble, W. J., & Wiley, D. E. (1974). Estimating content-acquiescence correlation by covariance structure analysis. Multivariate Behavioral Research, 9, 179-190.

Budescu, D. V., & Nevo, B. (1985). Optimal number of options: an investigation of the assumption of proportionality. Journal of Educational Measurement, 22, 183-196.

Carlton, S. T. (1987). Logical and critical thinking. Princeton, NJ: Educational Testing Service.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press

Costain, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: some empirical evidence for a mathematical proof. Educational and Psychological Measurement, 30, 353-358.

Costain, F. (1972). Three-choice versus four-choice items; Implications for reliability and validity of objective achievement tests. Educational and Psychological Measurement, 32, 1035-1038.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), Teaching thinking skills: Theory and practice (pp. 9-26). New York: Freeman & Co.

Frederiksen, N. (1983). Implications of theory for instruction in problem solving (ETS Research Rep. RR-83-19). Princeton, NJ: Educational Testing Service.

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. Psychological Bulletin, 105, 331-351.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. Journal of Educational Measurement, 12, 109-112.

Joreskog, K.G. (1970). A general method for analysis of covariance structures. Biometrika, 57, 239-251.

Kingston, N. M., & Dorans, N. J. (1982). The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory (GRE Board Professional Rep. No. 79-12bP, ETS Rep. 82-22). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1977). Optimal number of choices per item--a comparison of four approaches. Journal of Educational Measurement, 14, 33-38.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

McPeek, W. M., Chalifour, C., & Tucker, C. (1985, April). The analytical score: What it measures and what it doesn't. Paper presented at the meeting of the American Education Research Association, Chicago, IL.

Miller, R., & Wild, C. L. (Eds.) (1979). Restructuring the Graduate Record Examinations Aptitude Test (GRE Board Technical Rep.). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. Journal of Higher Education, 58, 658-682.

Powers, D. E., & Swinton, S. S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. Applied Psychological Measurement, 5(2), 141-158.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. Journal of Educational Psychology, 76, 266-278.

Raven, J. C. (1965). Progressive matrices. London: H. K. Lewis & Co., Ltd.

Rock, D. A., Bennett, R. E., and Jirele, T. (1988). Factor structure of the Graduate Record Examinations General Test in handicapped and nonhandicapped groups. Journal of Applied Psychology, 73, 383-392.

Ruch, G. M., & Stoddard, G. D. (1925). Comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 16, 89-103.

Schaeffer, G. A., & Kingston, N. M. (1988). Strength of the analytical factor of the GRE General Test in several subgroups: A full-information factor analysis approach (GRE Board Professional Rep. No. 86-7P). Princeton, NJ: Educational Testing Service.

Scriven, M. (1976). Reasoning, New York: McGraw-Hill.

Simon, H. A. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), Handbook of learning and cognitive processes (Vol. 5), Human information processing. Hillsdale, NJ: Erlbaum.

Stricker, L. J., & Rock, D. A. (1987). Factor structure of the GRE General Test in young and middle adulthood. Developmental Psychology, 23, 526-536.

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. Journal of Educational Psychology, 75, 104-115.

Swinton, S. S., Wild, C. L., & Wallmark, M. M. (1982). Investigation of practice effects on types of questions in the Graduate Record Examinations Aptitude Test (GRE Board Professional Rep. No. 80-1cP). Princeton, NJ: Educational Testing Service.

Tucker, C. (1985). Delineation of reasoning processes important to the construct validity of the Analytical Test (GRE Research Rep. No. 84-17). Princeton, NJ: Educational Testing Service.

Voss, J. F., Perkins, D., & Segal. J. (Eds.) (in press). Informal reasoning and education. Hillsdale, NJ: Erlbaum.

Ward, W. C., Carlson, S. B., & Woisetschlaeger, E. (1983). Ill-structured problems as multiple-choice items (GRE Board Professional Rep. No. 81-18P). Princeton, NJ: Education Testing Service.

Ward, W. C., Emmerich, W., Enright, M. K., Wightman, L., Powers, D., Gitomer, D., & Swinton, S. (1986, December). Validity of the GRE analytical measure: A program of research (GRE Proposal No. 86-19). Princeton, NJ: Educational Testing Service.

Werts, C. E., Linn, R. L., Joreskog, K. G. (1972). A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 32, 655-678.

Wild, C. L., Swinton, S. S., & Wallmark, M. M. (1982). Research leading to the revision of the format of the Graduate Record Examinations Aptitude Test in October 1981 (GRE Board Professional Rep. No. 80-1bP). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1982). A study of the validity of the restructured GRE Aptitude Test for predicting first-year performance in graduate study (GRE Board Research Rep. No. 78-6R, ETS Rep. No. 82-34). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1985). The relationship of GRE General Test item-type part scores to undergraduate grades (GRE Board Professional Rep. No. 81-22P, ETS Research Rep. No. 84-38). Princeton, NJ: Educational Testing Service.

Appendix A

Examples of Experimental Item Types

Analytical Reasoning

3-Option Multiple-Choice.....A-1

Multiple-Yes/No.....A-2

Logical Reasoning

3-Option Multiple-Choice.....A-3

Multiple-Yes/No.....A-4

Analysis of Explanations

3-Option Multiple-Choice.....A-5

Multiple-Yes/No.....A-11

Numerical Logical Reasoning

3-Option Multiple-Choice.....A-15

Multiple-Yes/No.....A-17

Contrasting Views

3-Option Multiple-Choice.....A-18

Pattern Identification

3-Option Multiple-Choice.....A-20

Section 1

Time Limit - 45 Minutes for Five Parts

Section 1, Part A: 9 Questions

Suggested Time - 10 minutes

Questions 1-5

An airline company is offering a particular group of people two package tours involving eight European cities—London, Madrid, Naples, Oslo, Paris, Rome, Stockholm, and Trieste. While half the group goes on tour 1 to visit five of the cities, the other half will go on tour 2 to visit the other three cities. The group must select the cities to be included in each tour. The selection must conform to the following restrictions:

Madrid cannot be in the same tour as Oslo.

Naples must be in the same tour as Rome.

If tour 1 includes Paris, it must also include London.

If tour 2 includes Stockholm, it cannot include Madrid.

1. Which of the following is an acceptable selection for the two tours? 1. (A) B C

	Tour 1	Tour 2
(A)	Madrid, Naples, Rome Stockholm, Trieste	Paris, London, Oslo
(B)	London, Madrid, Paris Rome, Trieste	Naples, Oslo, Stockholm
(C)	London, Madrid, Paris	Naples, Oslo, Rome Stockholm, Trieste

2. If tour 2 includes Rome, which of the following CANNOT be true? 2. A B (C)

(A) Trieste is in tour 1.

(B) Madrid is in tour 2.

(C) Stockholm is in tour 2.

GO ON TO THE NEXT PAGE.

123-132

An airline company is offering a particular group of people two package tours involving eight European cities--London, Madrid, Naples, Oslo, Paris, Rome, Stockholm, and Trieste. While half the group goes on tour 1 to visit five of the cities, the other half will go on tour 2 to visit the other three cities. The group must select the cities to be included in each tour. The selection must conform to the following restrictions:

Madrid cannot be in the same tour as Oslo.

Naples must be in the same tour as Rome.

If tour 1 includes Paris, it must also include London.

If tour 2 includes Stockholm, it cannot include Madrid.

If tour 2 includes Rome, can the following statement be true?

123. Trieste is in tour 1. 123. (Y) N

124. Madrid is in tour 2. 124. (Y) N

125. Stockholm is in tour 2. 125. Y (N)

If tour 2 includes Paris, must the following statement be true?

126. London is in tour 1. 126. Y (N)

127. Naples is in tour 1. 127. (Y) N

128. Stockholm is in tour 2. 128. Y (N)

GO ON TO THE NEXT PAGE.

6. Hittite tablets corroborate many of the descriptions of ancient life appearing in the Iliad and even list Greek cities that reportedly sent ships to Troy. This means that the Iliad is not creative literature, as has been believed, but history, and should be examined with the methods of historical science rather than literary criticism.

6. A (B) C

The author of the passage above makes which of the following assumptions?

I. A work should not be classified as creative literature if that work is known to record historical fact.

II. The Hittite tablets record actual events rather than nonfactual legends.

III. Cities and events mentioned in the Iliad but not in the Hittite tablets are fictitious.

(A) I only

(B) I and II only

(C) II and III only

7. A group of people saw a film of two cars colliding. Immediately afterward, half the group was asked questions about the cars "bumping" into one another, while the second half was asked the same questions with the verb "smash" substituted for "bump." In later descriptions of the collision, those in the second half were more likely to remember seeing broken glass.

7. A (B) C

The experiment described above best supports which of the following conclusions about eyewitness testimony?

(A) Most eyewitness testimony can be assumed to contain inaccurate elements.

(B) The manner in which a witness is questioned after an event can influence the recollection of the witness.

(C) A witness who is agitated at the time of an event is likely to give less accurate testimony than is a calm witness.

GO ON TO THE NEXT PAGE.

114-116

Hittite tablets corroborate many of the descriptions of ancient life appearing in the Iliad and even list Greek cities that reportedly sent ships to Troy. This means that the Iliad is not creative literature, as has been believed, but history, and should be examined with the methods of historical science rather than literary criticism.

Is the following an assumption made in the passage above?

114. A work should not be classified as creative literature if that work is known to record historical fact. 114. (Y) N

115. The Hittite tablets record actual events rather than nonfactual legends. 115. (Y) N

116. Cities and events mentioned in the Iliad but not in the Hittite tablets are fictitious. 116. Y (N)

117-119

A group of people saw a film of two cars colliding. Immediately afterward, half the group was asked questions about the cars "bumping" into one another, while the second half was asked the same questions with the verb "smash" substituted for "bump." In later descriptions of the collision, those in the second half were more likely to remember seeing broken glass.

Does the experiment described above support the following conclusion about eyewitness testimony?

117. Most eyewitness testimony can be assumed to contain inaccurate elements. 117. Y (N)

118. The manner in which a witness is questioned after an event can influence the recollection of the witness. 118. (Y) N

119. A witness who is agitated at the time of an event is likely to give less accurate testimony than is a calm witness. 119. Y (N)

GO ON TO THE NEXT PAGE.

Section 1, Part B: Instructions

Practice Time - 6 minutes

A situation and result will be presented, and questions will be asked about explaining the result. The following is an example:

Practice Questions Q1-Q4

Situation: The damming of the Palman River partially flooded the West Kenyan Wildlife Preserve and caused overcrowding of the animal population. Therefore, one hundred of the giraffes and one hundred of the Zimmerman gazelles were moved to the much larger East Kenyan preserve, where identical species of lions and giraffes as in the West Kenyan preserve and one species of gazelles, Allen gazelles, were already living. The only difference in climate was that the East Kenyan preserve averaged about ten inches less rain per year. In both preserves the prevailing winds were from the east and the terrain was mainly flat.

Result: After three years in the East Kenyan preserve, the Zimmerman gazelle population had diminished almost to the point of extinction.

In the context of the situation, the result needs explanation; you will be asked about explanations and statements relevant to explaining the result.

A statement is relevant to explaining the result if there is some possible adequate explanation of the result which the statement either supports or weakens.

Do not consider explanations that are remote and improbable. Borderline judgments about adequacy will not be required.

GO ON TO THE NEXT PAGE.

Practice Questions and Answers

Q1. Which of the following statements, if true, is relevant to some possible adequate explanation of the result?

(A) Zimmerman gazelles tend to panic and rush off frantically when hunted by lions.

(B) No zoo has succeeded in breeding Allen gazelles in captivity.

(C) Gazelles eat only grass.

Q2. Which of the following statements, if true, is relevant to some possible adequate explanation of the result?

(A) The Allen gazelles continued to flourish in the East Kenyan preserve.

(B) An earlier plan that was superseded did not include provisions for moving any giraffes to the East Kenyan preserve.

(C) The damming of the Palman River turned large areas of the West Kenyan preserve into a lake.

Q3. Which of the following statements, if true, is relevant to some possible adequate explanation of the result?

(A) The weather was normal in East Kenya during the three years after the transfer.

(B) Kenya's efforts to increase hydroelectric power caused the overcrowding in the West Kenyan preserve.

(C) The species of Zimmerman gazelles is not in danger of extinction because many zoos throughout the world contain populations of the species.

Q4. Which of the following, if true, CANNOT provide the basis for an adequate explanation of the result?

(A) The Kenyan government was warned before the animal transfer that such transfers were frequently unsuccessful.

(B) The Zimmerman gazelles contracted a disease that was new to them from the Allen gazelles and succumbed to it.

(C) The Zimmerman gazelle is famous for its delicious meat, and there was far more poaching in the East Kenyan preserve than there was in the West Kenyan preserve.

GO ON TO THE NEXT PAGE.

Answers for Practice Questions Q1–Q4:

Q1. C (A) is irrelevant, because there had already been lions of the same species in the West Kenyan preserve where the Zimmerman gazelles had previously flourished, so the lions were not new to the Zimmerman gazelles.

(B) is irrelevant, because it was Zimmerman gazelles, not Allen gazelles, that suffered the population decline, and because the captivity of the Zimmerman gazelles was not permanent, as in a zoo, but temporary, for the purposes of transportation.

(C) strengthens, and so is relevant to, a possible explanation that the Zimmerman gazelles did not receive proper nutrition in the East Kenyan preserve, because the Allen gazelles were more efficient in cropping the short grass that grew there, and little was left for the Zimmerman gazelles.

Q2. A (A) weakens, and so is relevant to, a possible explanation that a severe and protracted drought in the East Kenyan preserve made it almost impossible for any kind of gazelle to flourish there.

Q3. A (A) weakens, and so is relevant to, the possible explanation cited for Q2 above.

Q4. A (A) does not adequately explain the result, because no reason for the lack of success is given.

(B) provides the basis for an adequate explanation of the result, because the opportunity for the Zimmerman gazelles to contract the fatal disease resulted from the transfer.

(C) provides the basis for an adequate explanation of the result, because the selective poaching could decimate the population of Zimmerman gazelles.

Section 1, Part C: 8 Questions

Suggested Time - 11 minutes

Questions 10-13

Situation: At least once each summer during the ten years since their house had been built, Thelma and Raymond Ashe discovered an inch or two of water in their basement after severe storms; they also found that one wall of the basement was damp. Although the damage was never serious, the Ashes worried about the potential for damage if a major storm should lead to more severe flooding. Therefore, they had a waterproofing compound applied to the cement walls and floor of the basement, with extra attention to cracks and holes. Then they had extra concrete added to the outer walls of the foundation. Finally, they had a trench dug around the house and out from it to carry water away from the foundation.

Result: The following year, during a torrential rain lasting three days, even though the basement remained dry, the house and basement shifted, almost causing the house to collapse.

In the context of the situation, the result needs explanation; you will be asked about explanations and statements relevant to explaining the result.

A statement is relevant to explaining the result if there is some possible adequate explanation of the result which the statement either supports or weakens.

Do not consider explanations that are remote and improbable. Borderline judgments about adequacy will not be required.

GO ON TO THE NEXT PAGE.

10. Which of the following statements, if true, is relevant to some possible adequate explanation of the result?

    (A) The weather bureau, in predicting the storm, had underestimated its severity.

    (B) The trench was not lined with material such as stones, tile, or concrete.

    (C) The three-day rain caused severe flood damage in surrounding communities, in areas other than the place where the Ashes lived.

10. A (B) C

11. Which of the following statements, if true, is relevant to some possible adequate explanation of the result?

    (A) The soil in which the basement was built was sandy.

    (B) After about ten years, the waterproofing compound deteriorates and must then be re-applied.

    (C) The roof of the house was made of pieces of slate.

11. (A) B C

12. Which of the following, if true, could provide the basis for an adequate explanation of the result?

    (A) After the Ashes applied waterproofing to the basement, they finished the walls with paint and paneling.

    (B) A building permit for the modifications to the basement was issued to the Ashes on the basis of an inspector's certification that the modifications would not impair the structural soundness of the house.

    (C) Water carried away from the house by the trench eroded a portion of a hillside just below the house and caused a landslide.

12. A B (C)

GO ON TO THE NEXT PAGE.

Section 1, Part B: Instructions

Practice Time - 6 minutes

A situation and result will be presented, and questions will be asked about explaining the result. The following is an example:

Practice Statements S1-S4

Situation: The damming of the Palman River partially flooded the West Kenyan Wildlife Preserve and caused overcrowding of the animal population. Therefore, one hundred of the giraffes and one hundred of the Zimmerman gazelles were moved to the much larger East Kenyan preserve, where identical species of lions and giraffes as in the West Kenyan preserve and one species of gazelles, Allen gazelles, were already living. The only difference in climate was that the East Kenyan preserve averaged about ten inches less rain per year. In both preserves the prevailing winds were from the east and the terrain was mainly flat.

Result: After three years in the East Kenyan preserve, the Zimmerman gazelle population had diminished almost to the point of extinction.

In the context of the situation, the result needs explanation; you will be asked about explanations and statements relevant to explaining the result.

A statement is relevant to explaining the result if there is some possible adequate explanation of the result which the statement either supports or weakens.

Do not consider explanations that are remote and improbable. Borderline judgments about adequacy will not be required.

GO ON TO THE NEXT PAGE.

Practice Statements and Answers

Is the following statement, if true, relevant to an explanation of the result?

S1. No zoo has succeeded in breeding Allen gazelles in captivity.

S2. The weather was normal in East Kenya during the three years after the transfer.

Could the following, if true, provide the basis for an adequate explanation of the result?

S3. The animals successfully rounded up for the transfer included primarily the weaker Zimmerman gazelles, which then lost out in competition for grass with the Allen gazelles.

S4. Kenya's efforts to increase hydroelectric power caused the overcrowding in the West Kenyan preserve.

Answers to practice statements S1-S4:

S1. N This statement is irrelevant, because it was Zimmerman gazelles, not Allen gazelles, that suffered the population decline, and because the captivity of the Zimmerman gazelles was not permanent, as in a zoo, but temporary, for the purposes of transportation.

S2. Y This statement weakens, and so is relevant to, a possible explanation that there was a drought in the East Kenyan preserve so severe and protracted that no species of gazelle was able to survive there.

S3. Y This statement adequately explains the result, because it shows how the process of transfer worked to diminish the vitality of the stock of transferred Zimmerman gazelles, and how competition with the Allen gazelles was then sufficient to produce the result.

S4. N This statement does not adequately explain the result; rather, it explains the damming of the river, which is only a part of the situation leading up to the result. There is still an unexplained gap between the situation and the result, namely, the reason why the Zimmerman gazelles failed to flourish in the new location.

GO ON TO THE NEXT PAGE WHEN YOU ARE READY.

96-104

Situation: At least once each summer during the ten years since their house had been built, Thelma and Raymond Ashe discovered an inch or two of water in their basement after severe storms; they also found that one wall of the basement was damp. Although the damage was never serious, the Ashes worried about the potential for damage if a major storm should lead to more severe flooding. Therefore, they had a waterproofing compound applied to the cement walls and floor of the basement, with extra attention to cracks and holes. Then they had extra concrete added to the outer walls of the foundation. Finally, they had a trench dug around the house and out from it to carry water away from the foundation.

Result: The following year during a torrential rain lasting three days, even though the basement remained dry, the house and basement shifted, almost causing the house to collapse.

In the context of the situation, the result needs explanation; you will be asked about explanations and statements relevant to explaining the result.

A statement is relevant to explaining the result if there is some possible adequate explanation of the result which the statement either supports or weakens.

Do not consider explanations that are remote and improbable. Borderline judgments about adequacy will not be required.

GO ON TO THE NEXT PAGE.

Is the following statement, if true, relevant to an explanation of the result?

96. The weather bureau, in predicting the storm, had underestimated its severity. 96. Y (N)

97. The trench was not lined with material such as stones, tile, or concrete. 97. (Y) N

98. The three-day rain caused severe flood damage in surrounding communities, in areas other than the place where the Ashes lived. 98. Y (N)

99. Either the waterproofing or the extra concrete sealed the place or places where water had previously entered the basement. 99. (Y) N

100. The soil in which the basement was built was sandy. 100. (Y) N

Could the following, if true, provide the basis for an adequate explanation of the result?

101. In the northwest corner of the house, the basement was set directly on bedrock. 101. Y (N)

102. Water saturated the soil below the sealed basement and allowed the house to act like an unstable boat. 102. (Y) N

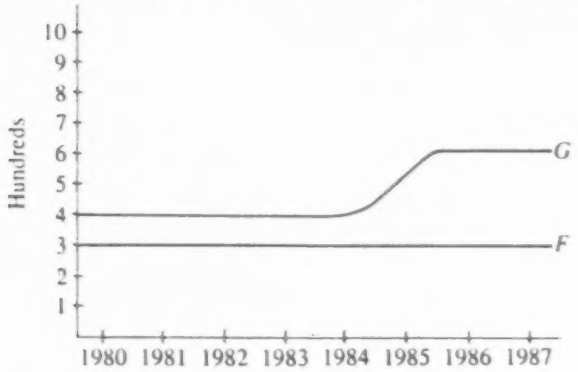
103. Water carried away from the house by the trench eroded a portion of a hillside just below the house and caused a landslide. 103. (Y) N

104. A building permit for the modifications to the basement was issued to the Ashes on the basis of an inspector's certification that the modifications would not impair the structural soundness of the house. 104. Y (N)

GO ON TO THE NEXT PAGE.

Questions 39-40 are based on the following graph.

POPULATION OF BIRD SPECIES *F* AND *G*
IN THE STATE OF AVIARYA FROM 1980-1987



39. Which of the following, if true, could help explain the data illustrated in the graph on differences in population totals for species F and G?

39. A B C

(A) In 1984, harsh winter weather caused an unusually large portion of the species-F population temporarily to migrate south of Aviarya.

(B) In 1984, males of species G outnumbered females of species G for the first time since 1981.

(C) In 1984, species G was afforded protected status as the state bird of Aviarya.

GO ON TO THE NEXT PAGE.

40. A pesticide that had no direct harmful effect on birds of either species F or species G was widely used in Aviarya before 1984. It is claimed that the change in population patterns illustrated above occurred because the use of the pesticide was discontinued in 1984. Each of the following, if true, weakens this claim EXCEPT:

(A) In 1984, a means of controlling a disease that makes eggshells of birds of species G fragile was introduced in Aviarya.

(B) In 1984, the results of a study aimed at assessing the effects of the pesticide were published.

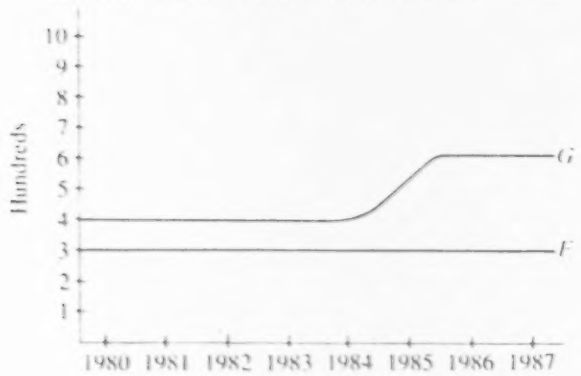
(C) In 1984, the pesticide was proved very effective in selectively controlling the predators of insects that are the preferred food of species G.

40. A (B) C

GO ON TO THE NEXT PAGE.

78-81 are based on the following graph.

POPULATION OF BIRD SPECIES *F* AND *G*
IN THE STATE OF AVIARYA FROM 1980-1987



Could the following, if true, be a factor explaining, at least in part, the data illustrated in the graph on differences in population totals for species F and G?

78. In 1984, harsh winter weather caused an unusually large portion of the species-F population temporarily to migrate south of Aviarya. 78. Y (N)

79. In 1984, species G was afforded protected status as the state bird of Aviarya. 79. (Y) N

A pesticide that had no direct harmful effect on birds of either species F or species G was widely used in Aviarya before 1984. It is claimed that the change in population patterns illustrated above occurred because the use of the pesticide was discontinued in 1984. Does the following, if true, weaken this claim?

80. In 1984, a means of controlling a disease that makes eggshells of birds of species G fragile was introduced in Aviarya. 80. (Y) N

81. In 1984, the results of a study aimed at assessing the effects of the pesticide were published. 81. Y (N)

GO ON TO THE NEXT PAGE.

Questions 30–34 are based on the following contrasting views.

View I: A painting's form--its use of line, color, and shape—arouses the viewer's aesthetic sense, whereas its content, if appealing or interesting, often interferes with the viewer's aesthetic appreciation. Abstract masterpieces lacking discernible subjects, because they provide a source of pure aesthetic experience, as opposed to sentimental or intellectual experience, are the highest form of art.

View II: Art engages the mind, inspires the soul, and arouses the senses. In great art, form and content cooperate perfectly, so that the eye, stimulated by the formal beauties of line, color, and shape, lingers to search out the deeper truth of what it sees. Aesthetic experience satisfies so deeply precisely because it involves all our faculties, sensory, intellectual, and spiritual.

30. Which of the following, if true, would provide a basis for criticizing view I, but not necessarily view II? 30. A B (C)

(A) Since the response of a viewer to a work of art can become dulled by familiarity, not all viewers respond in the same way to great art.

(B) All objects possessing aesthetic value, whether they are art objects, functional implements, or things found in nature, share one common characteristic: a pleasing form.

(C) All complex mental processes have important intellectual or emotional components, since even the simplest act of perception is affected by memory, expectation, and desire.

31. The two views are in agreement with regard to which of the following? 31. (A) B C

(A) Which elements in a work constitute its form

(B) How a work's form interacts with its content

(C) Which works provide the best source of aesthetic experience

GO ON TO THE NEXT PAGE.

32. The two viewpoints differ most in the degree to which they hold that great works of art 32. A (B) C

(A) provide the best source of aesthetic experience

(B) are great, in part, because they are intellectually interesting

(C) rely on the use of line, shape, and form to stimulate viewers

33. The acceptance of view II, in contrast to view I, requires acceptance of the view that 33. (A) B C

(A) an aesthetic experience involving the use of multiple faculties is more satisfying than one involving only one faculty

(B) paintings lacking beauty of form can still be great works of art if their content is intellectually or spiritually challenging

(C) human beings possess an aesthetic sense that is separate from their other mental faculties

34. Which of the following statements is inconsistent with (cannot be true along with) each view presented above? 34. A B (C)

(A) Art serves aesthetic ends, so, in a work of art, form should be primary and arresting, and content should be restricted or absent.

(B) Art involves the total person, so, in a work of art, form and content are equally important and should be coequal partners.

(C) Art speaks to everyone in a society, so, in a work of art, content should be dominant and immediately accessible, and form should be kept simple.

GO ON TO THE NEXT PAGE.

Section 1, Part D: Instructions

Practice Time - 6 minutes

INSTRUCTIONS FOR QUESTIONS ON NUMBER SERIES

In this test, a number series is composed of exactly seven whole numbers (positive integers).

Example: 2, 4, 6, 8, 10, 12, 14

Each number in a series except the first (leftmost) is calculated from the number preceding (to the immediate left) by applying a series rule. For the example above an applicable series rule is the following: "The numbers successively increase by 2." This series rule can be represented as follows:

2, 4, 6, 8, 10, 12, 14
(+2) (+2) (+2) (+2) (+2) (+2)

A different example of a number series is the following:

Example: 1, 3, 7, 15, 31, 63, 127
(x2,+1) (x2,+1) (x2,+1) (x2,+1) (x2,+1) (x2,+1)

For this example an applicable series rule is the following: "The number is calculated by multiplying the preceding number by 2 and then adding 1 to the product." This calculation is represented in the example as (x2,+1).

The number must be calculable by performing either one arithmetic operation, as in the first example above, or two arithmetic operations, as in the second example above. An "arithmetic operation" is limited to adding 1,2, or 3, subtracting 1,2, or 3, multiplying by 2 or 3, and dividing by 2 or 3. If two arithmetic operations are performed, they cannot both be addition, or both be subtraction, or both be multiplication, or both be division.

In each example thus far a single formula has been used to calculate the numbers. However, an applicable series rule may utilize more than one formula, in which case the series rule must conform to one of the four patterns described below.

GO ON TO THE NEXT PAGE.

1 1

Pattern: The second, fourth, and sixth numbers are each calculated one way, and the third, fifth, and seventh numbers are each calculated another way.

Example: 2, 4, 3, 6, 5, 10, 9
(x2) (-1) (x2) (-1) (x2) (-1)

For this example an applicable series rule is the following: "The second, fourth, and sixth numbers are each calculated by doubling the preceding number; the third, fifth, and seventh numbers are each calculated by subtracting 1 from the preceding number."

Pattern: The second and fifth numbers are each calculated one way, the third and sixth numbers are each calculated another way, and the fourth and seventh numbers are each calculated still another way.

Example: 1, 4, 8, 5, 8, 16, 9
(+3) (x2) (÷2,+1) (+3) (x2) (÷2,+1)

For this example an applicable series rule is the following: "The second and fifth numbers are each calculated by adding 3 to the preceding number; the third and sixth numbers are each calculated by doubling the preceding number; the fourth and seventh numbers are each calculated by dividing the preceding number by 2 and then adding 1 to the result."

Pattern: The second, third, and fourth numbers are each calculated one way, and the fifth, sixth, and seventh numbers are each calculated another way.

Example: 5, 8, 11, 14, 11, 8, 5
(+3) (+3) (+3) (-3) (-3) (-3)

For this example an applicable series rule is the following: "The second, third, and fourth numbers are each calculated by adding 3 to the preceding number; the fifth, sixth, and seventh numbers are each calculated by subtracting 3 from the preceding number."

GO ON TO THE NEXT PAGE.

Pattern: The second and third numbers are each calculated one way, the fourth and fifth numbers are each calculated another way, and the sixth and seventh numbers are each calculated still another way.

Example:

3,	2,	1,	3,	7,	3,	1
	(-1)	(-1)	(x2,+1)	(x2,+1)	(-1,÷2)	(-1,÷2)

For this example an applicable series rule is the following: "The second and third numbers are each calculated by subtracting 1 from the preceding number; the fourth and fifth numbers are each calculated by multiplying the preceding number by 2 and then adding 1 to the product; the sixth and seventh numbers are each calculated by subtracting 1 from the preceding number and then dividing the remainder by 2."

GO ON TO THE NEXT PAGE WHEN YOU ARE READY.

Section 1, Part E: 8 Questions

Suggested Time - 12 minutes

You may refer back to the instructions for Number Series at any time.

Next to each question below a series is presented, followed by three options, (A), (B), and (C), each of which is a series. Select the option for which an accurate and complete series rule is also an accurate and complete series rule for the series presented next to the question.

Questions 18-25

18.		4,	3,	6,	8,	7,	14,	16	18. (A) B C
	(A)	2,	1,	2,	4,	3,	6,	8	
	(B)	3,	2,	4,	6,	4,	8,	10	
	(C)	5,	4,	7,	11,	10,	13,	23	

19.		2,	4,	3,	6,	5,	10,	9	19. (A) B C
	(A)	1,	2,	1,	2,	1,	2,	1	
	(B)	3,	6,	4,	8,	6,	12,	10	
	(C)	5,	7,	6,	8,	7,	9,	8	

GO ON TO THE NEXT PAGE.

Appendix B

Participating Test Centers

Participating Test Centers

Archbishop Carroll High School, Washington, DC
California State University, Fresno, CA
Dallas Baptist University, Dallas TX
Douglass College, New Brunswick, NJ
El Camino Real High School, Woodland Hills, CA
Emory University, Atlanta, GA
Jackson State University, Jackson, MS
Montana State University, Bozeman, MT
New York University-Trinity Place, New York, NY
Ohio State University, Columbus, OH
Rutgers University, Newark, NJ
Simmons College, Boston, MA
South Dakota State University, Brookings, SD
Southeastern Oklahoma State College, Durant, OK
Temple University, Philadelphia, PA
Texas A & M University, College Station, TX
University of Arizona, Tucson, AZ
University of Colorado, Denver, CO
University of Detroit, Detroit, MI
University of Florida, Gainesville, FL
University of Hawaii, Honolulu, HI
University of Illinois, Urbana, IL
University of Kansas, Lawrence, KS
University of Louisville, Louisville, KY
University of Michigan, Ann Arbor, MI
University of Nevada, Reno, NV
University of North Carolina, Greensboro, NC
University of North Texas, Denton, TX
University of Northern Iowa, Cedar Falls, IA
University of Pittsburgh, Pittsburgh, PA
University of South Florida, Tampa, FL
University of Southern Maine, Portland, ME
University of Washington, Seattle, WA
University of Wisconsin, Madison, WI
West Chester University of Pennsylvania, West Chester, PA

Appendix C
Correlation Matrices: Observed Correlations and Correlations
Corrected for Unreliability

CORRELATION MATRIX FORM 2B

	AR2	LR2	AX2	NLR2	QUANT	REG MATH	DATA INT	ANALREAS	LOGLREAS	SENTCOMP	READCOMP	ANALOGY
AR2	1.0000	0.5143	0.4070	0.4464	0.5450	0.4675	0.4195	0.5259	0.3657	0.3679	0.3939	0.2752
LR2	0.5143	1.0000	0.5027	0.5863	0.5014	0.4467	0.4011	0.4586	0.5511	0.5694	0.5652	0.4644
AX2	0.4070	0.5027	1.0000	0.5273	0.3469	0.2986	0.3382	0.4276	0.4368	0.4572	0.4784	0.3921
NLR2	0.4464	0.5863	0.5273	1.0000	0.5232	0.5075	0.5110	0.5130	0.5653	0.5466	0.5508	0.4650
QUANT	0.5450	0.5014	0.3469	0.5232	1.0000	0.7903	0.6297	0.6529	0.4787	0.4256	0.4676	0.3116
REG MATH	0.4675	0.4467	0.2986	0.5075	0.7903	1.0000	0.6304	0.6349	0.4629	0.3953	0.4182	0.2794
DATA INT	0.4195	0.4011	0.3382	0.5110	0.6297	0.6304	1.0000	0.5083	0.3890	0.3861	0.4566	0.2729
ANALREAS	0.5259	0.4586	0.4276	0.5130	0.6529	0.6349	0.5083	1.0000	0.4785	0.4452	0.5416	0.3603
LOGLREAS	0.3657	0.5511	0.4368	0.5653	0.4787	0.4629	0.3890	0.4785	1.0000	0.5721	0.6356	0.4977
SENTCOMP	0.3679	0.5694	0.4572	0.5466	0.4256	0.3953	0.3861	0.4452	0.5721	1.0000	0.6658	0.6260
READCOMP	0.3939	0.5652	0.4784	0.5508	0.4676	0.4182	0.4566	0.5416	0.6356	0.6658	1.0000	0.5420
ANALOGY	0.2752	0.4644	0.3921	0.4650	0.3116	0.2794	0.2729	0.3603	0.4977	0.6260	0.5420	1.0000
ANTONYM	0.2958	0.5047	0.4087	0.4974	0.3700	0.3088	0.3346	0.3802	0.5537	0.6663	0.6298	0.6576
VRB RGHT	0.3927	0.6182	0.5103	0.6052	0.4673	0.4133	0.4327	0.5118	0.6687	0.8495	0.8547	0.7958
QNT RGHT	0.5463	0.5121	0.3625	0.5689	0.9482	0.9208	0.7698	0.6845	0.5042	0.4501	0.4948	0.3237
AN RGHT	0.5416	0.5392	0.4805	0.5884	0.6796	0.6601	0.5337	0.9675	0.6849	0.5341	0.6323	0.4422
AN CONV	0.5602	0.5560	0.4932	0.5994	0.6472	0.6366	0.5077	0.9485	0.6660	0.5088	0.6098	0.4052
QNT CONV	0.5709	0.5313	0.3808	0.5784	0.9111	0.8937	0.7354	0.6492	0.4769	0.4082	0.4583	0.2687
VRB CONV	0.4048	0.6322	0.5203	0.6091	0.4333	0.3880	0.4036	0.4849	0.6518	0.8262	0.8381	0.7736
GPA	0.2459	0.2272	0.3243	0.3095	0.3116	0.3095	0.2741	0.2831	0.2738	0.2783	0.3097	0.2081
SEX	-0.1251	-0.1670	-0.0333	-0.2406	-0.3439	-0.3592	-0.3320	-0.1533	-0.1884	-0.1625	-0.1455	-0.0992

	ANTONYM	VRB RGHT	QNT RGHT	AN RGHT	AN CONV	QNT CONV	VRB CONV	GPA	SEX
AR2	0.2958	0.3927	0.5463	0.5416	0.5602	0.5709	0.4048	0.2459	-0.1251
LR2	0.5047	0.6182	0.5121	0.5392	0.5560	0.5313	0.6322	0.2272	-0.1670
AX2	0.4087	0.5103	0.3625	0.4805	0.4932	0.3808	0.5203	0.3243	-0.0333
NLR2	0.4974	0.6052	0.5689	0.5884	0.5994	0.5784	0.6091	0.3095	-0.2406
QUANT	0.3700	0.4673	0.9482	0.6796	0.6472	0.9111	0.4333	0.3116	-0.3439
REG MATH	0.3088	0.4133	0.9208	0.6601	0.6366	0.8937	0.3880	0.3095	-0.3592
DATA INT	0.3346	0.4327	0.7698	0.5337	0.5077	0.7354	0.4036	0.2741	-0.3320
ANALREAS	0.3802	0.5118	0.6845	0.9675	0.9485	0.6492	0.4849	0.2831	-0.1533
LOGLREAS	0.5537	0.6687	0.5042	0.6849	0.6660	0.4769	0.6518	0.2738	-0.1884
SENTCOMP	0.6663	0.8495	0.4501	0.5341	0.5088	0.4082	0.8262	0.2783	-0.1625
READCOMP	0.6298	0.8547	0.4948	0.6323	0.6098	0.4583	0.8381	0.3097	-0.1455
ANALOGY	0.6576	0.7958	0.3237	0.4422	0.4052	0.2687	0.7736	0.2081	-0.0992
ANTONYM	1.0000	0.8906	0.3777	0.4748	0.4525	0.3435	0.8802	0.2716	-0.1629
VRB RGHT	0.8906	1.0000	0.4880	0.6171	0.5871	0.4409	0.9801	0.3202	-0.1709
QNT RGHT	0.3777	0.4880	1.0000	0.7131	0.6821	0.9634	0.4548	0.3363	-0.3831
AN RGHT	0.4748	0.6171	0.7131	1.0000	0.9788	0.6760	0.5900	0.3142	-0.1815
AN CONV	0.4525	0.5871	0.6821	0.9788	1.0000	0.6986	0.5964	0.3098	-0.1704
QNT CONV	0.3435	0.4409	0.9634	0.6760	0.6986	1.0000	0.4572	0.3311	-0.3796
VRB CONV	0.8802	0.9801	0.4548	0.5900	0.5964	0.4572	1.0000	0.3186	-0.1660
GPA	0.2716	0.3202	0.3363	0.3142	0.3098	0.3311	0.3186	1.0000	-0.0127
SEX	-0.1629	-0.1709	-0.3831	-0.1815	-0.1704	-0.3796	-0.1660	-0.0127	1.0000

BEST COPY AVAILABLE

CORRELATIONS CORRECTED FOR ATTENUATION FORM 2B MULTIPLE-YES/NO

	AR2	LR2	AX2	NLR2	QUANT	REG MATH	DATA INT	ANALREAS	LOGLREAS	SENTCOMP	READCOMP	ANALOGY
AR2	1.00	0.72	0.59	0.59	0.68	0.60	0.59	0.64	0.53	0 50	0.51	0.39
LR2	0.72	1.00	0.78	0.84	0.67	0.62	0.61	0.60	0.86	0.84	0.79	0.72
AX2	0.59	0.78	1.00	0.78	0.48	0.43	0.53	0.58	0.70	0.70	0.69	0.63
NLR2	0.59	0.84	0.78	1.00	0.66	0.66	0.73	0.64	0.83	0.76	0.73	0.68
QUANT	0.68	0.67	0.48	0.66	1.00	0.97	0.85	0.76	0.66	0.56	0.58	0.43
REG MATH	0.60	0.62	0.43	0.66	0.97	1.00	0.88	0.76	0.66	0.53	0.54	0.40
DATA INT	0.59	0.61	0.53	0.73	0.85	0.88	1.00	0.67	0.61	0.57	0.64	0.42
ANALREAS	0.64	0.60	0.58	0.64	0.76	0.76	0.67	1.00	0.65	0.57	0.66	0.48
LOGLREAS	0.53	0.86	0.70	0.83	0.66	0.66	0.61	0.65	1.00	0.87	0.92	0.79
SENTCOMP	0.50	0.84	0.70	0.76	0.56	0.53	0.57	0.57	0.87	1.00	0.91	0.94
READCOMP	0.51	0.79	0.69	0.73	0.58	0.54	0.64	0.66	0.92	0.91	1.00	0.78
ANALOGY	0.39	0.72	0.63	0.68	0.43	0.40	0.42	0.48	0.79	0.94	0.78	1.00
ANTONYM	0.37	0.67	0.56	0.62	0.44	0.38	0.45	0.44	0.92	0.78	0.78	0.90
VRB RGHT	0.47	0.80	0.68	0.74	0.53	0.49	0.56	0.57	0.79	0.87	0.78	1.05
QNT RGHT	0.65	0.66	0.48	0.69	1.09	1.09	1.00	0.77	0.76	1.07	1.02	0.43
AN RGHT	0.66	0.71	0.65	0.73	0.79	0.79	0.70	1.10	0.88	0.57	0.59	0.59
AN CONV	0.64	0.68	0.63	0.70	0.71	0.72	0.63	1.01	0.67	0.68	0.77	0.51
QNT CONV	0.65	0.65	0.48	0.67	1.00	1.01	0.91	0.69	0.92	0.61	0.70	0.34
VRB CONV	0.46	0.78	0.66	0.71	0.47	0.44	0.50	0.52	0.82	0.49	0.52	0.97
GPA	0.28	0.28	0.41	0.36	0.34	0.35	0.34	0.30	0.35	0.99	0.96	0.26
SEX	-0.14	-0.21	-0.04	-0.28	-0.38	-0.40	-0.41	-0.16	-0.24	0.33	0.35	0.97
										-0.19	-0.17	-0.12

	ANTONYM	VRB RGHT	QNT RGHT	AN RGHT	AN CONV	QNT CONV	VRB CONV	GPA	SEX
AR2	0.37	0.47	0.65	0.66	0.64	0.65	0.46	0.28	-0.14
LR2	0.67	0.80	0.66	0.71	0.68	0.65	0.78	0.28	-0.21
AX2	0.56	0.68	0.48	0.65	0.63	0.48	0.66	0.41	-0.04
NLR2	0.62	0.74	0.69	0.73	0.70	0.67	0.71	0.36	-0.28
QUANT	0.44	0.53	1.09	0.79	0.71	1.00	0.47	0.34	-0.38
REG MATH	0.38	0.49	1.09	0.79	0.72	1.01	0.44	0.35	-0.40
DATA INT	0.45	0.56	1.00	0.70	0.63	0.91	0.50	0.34	-0.41
ANALREAS	0.44	0.57	0.77	1.10	1.01	0.69	0.52	0.34	-0.41
LOGLREAS	0.76	0.88	0.67	0.92	0.84	0.60	0.82	0.30	-0.16
SENTCOMP	0.87	1.07	0.57	0.68	0.61	0.49	0.99	0.35	-0.24
READCOMP	0.78	1.02	0.59	0.77	0.70	0.52	0.96	0.33	-0.19
ANALOGY	0.90	1.05	0.43	0.59	0.51	0.34	0.97	0.35	-0.17
ANTONYM	1.00	1.01	0.43	0.55	0.49	0.37	0.95	0.26	-0.12
VRB RGHT	1.01	1.00	0.54	0.69	0.61	0.46	1.03	0.29	-0.18
QNT RGHT	0.43	0.54	1.00	0.80	0.71	1.01	0.48	0.34	-0.18
AN RGHT	0.55	0.69	0.80	1.00	1.04	0.72	0.63	0.35	-0.40
AN CONV	0.49	0.61	0.71	1.04	1.00	0.70	0.60	0.33	-0.19
QNT CONV	0.37	0.46	1.01	0.72	0.70	1.00	0.46	0.31	-0.17
VRB CONV	0.95	1.03	0.48	0.63	0.60	0.46	1.00	0.32	-0.17
GPA	0.29	0.34	0.35	0.33	0.31	0.33	0.32	1.00	-0.01
SEX	-0.18	-0.18	-0.40	-0.19	-0.17	-0.38	-0.17	-0.01	1.00

CORRELATION MATRIX FORM 3A

	AR3	LR3	CV3	AX3	NLR3	PI3	QUANT	REG MATH	DATA INT	ANALREAS	LOGLREAS	SENTCOMP
AR3	1.0000	0.5716	0.3923	0.3752	0.5041	0.5374	0.6380	0.5989	0.5045	0.6840	0.4929	0.4122
LR3	0.5716	1.0000	0.5598	0.5131	0.6545	0.4860	0.5726	0.5112	0.4531	0.5134	0.6103	0.5593
CV3	0.3923	0.5598	1.0000	0.5166	0.5741	0.3822	0.4451	0.3988	0.3465	0.4467	0.5608	0.6052
AX3	0.3752	0.5131	0.5166	1.0000	0.4851	0.4039	0.3697	0.3189	0.2384	0.3647	0.4588	0.5223
NLR3	0.5041	0.6545	0.5741	0.4851	1.0000	0.5239	0.5641	0.4953	0.4980	0.5025	0.5511	0.5131
PI3	0.5374	0.4860	0.3822	0.4039	0.5239	1.0000	0.5300	0.5134	0.3978	0.5501	0.4607	0.4029
QUANT	0.6380	0.5726	0.4451	0.3697	0.5641	0.5300	1.0000	0.7649	0.5832	0.6895	0.4939	0.4734
REG MATH	0.5989	0.5112	0.3988	0.3189	0.4953	0.5134	0.7649	1.0000	0.6369	0.6736	0.4587	0.4599
DATA INT	0.5045	0.4531	0.3465	0.2384	0.4980	0.3978	0.5832	0.6369	1.0000	0.5467	0.4409	0.4228
ANALREAS	0.6840	0.5134	0.4467	0.3647	0.5025	0.5501	0.6895	0.6736	0.5467	1.0000	0.5136	0.4452
LOGLREAS	0.4929	0.6103	0.5608	0.4588	0.5511	0.4607	0.4939	0.4587	0.4409	0.5136	1.0000	0.5669
SENTCOMP	0.4122	0.5593	0.6052	0.5223	0.5131	0.4029	0.4734	0.4599	0.4228	0.4452	0.5669	1.0000
READCOMP	0.5416	0.6208	0.6212	0.5307	0.6032	0.4723	0.5491	0.4667	0.3929	0.5401	0.6103	0.6531
ANALOGY	0.4217	0.5402	0.5897	0.4841	0.4785	0.3178	0.4092	0.3574	0.3664	0.3886	0.5644	0.6467
ANTONYM	0.4057	0.5087	0.5848	0.4261	0.4224	0.2809	0.4117	0.3682	0.3602	0.3847	0.5415	0.6510
VRB RGHT	0.5263	0.6522	0.7017	0.5695	0.5902	0.4322	0.5434	0.4845	0.4486	0.5192	0.6674	0.8447
QNT RGHT	0.6682	0.5892	0.4575	0.3654	0.5871	0.5574	0.9371	0.9207	0.7514	0.7330	0.5251	0.5115
AN RGHT	0.7036	0.5941	0.5252	0.4290	0.5688	0.5835	0.7085	0.6856	0.5753	0.9718	0.7014	0.5256
AN CONV	0.7012	0.5930	0.5256	0.4273	0.5687	0.5818	0.7068	0.6849	0.5716	0.9707	0.6989	0.5220
QNT CONV	0.6690	0.5876	0.4584	0.3651	0.5872	0.5568	0.9359	0.9208	0.7513	0.7339	0.5245	0.5125
VRB CONV	0.5212	0.6422	0.6983	0.5578	0.5863	0.4256	0.5330	0.4771	0.4463	0.5094	0.6617	0.8311
GPA	0.2437	0.2704	0.2724	0.2930	0.2925	0.2920	0.2838	0.2756	0.1899	0.3250	0.5216	0.2995
SEX	-0.1083	-0.2142	-0.1244	-0.0385	-0.1632	-0.1197	-0.2589	-0.2926	-0.2765	-0.0553	-0.1716	-0.1023

CORRELATION MATRIX FORM 3A

	READCOMP	ANALOGY	ANTONYM	VRB RGHT	QNT RGHT	AN RGHT	AN CONV	QNT CONV	VRB CONV	GPA	SEX
AR3	0.5416	0.4217	0.4057	0.5263	0.6682	0.7036	0.7012	0.6690	0.5212	0.2437	-0.1083
LR3	0.6208	0.5402	0.5087	0.6522	0.5892	0.5941	0.5930	0.5876	0.6422	0.2704	-0.2142
CV3	0.6212	0.5897	0.5848	0.7017	0.4575	0.5252	0.5256	0.4584	0.6983	0.2724	-0.1244
AX3	0.5307	0.4841	0.4261	0.5695	0.3654	0.4290	0.4273	0.3651	0.5578	0.2930	-0.0385
NLR3	0.6032	0.4785	0.4224	0.5902	0.5871	0.5688	0.5687	0.5872	0.5863	0.2925	-0.1632
PI3	0.4723	0.3178	0.2809	0.4322	0.5574	0.5835	0.5818	0.5568	0.4256	0.2920	-0.1197
QUANT	0.5491	0.4092	0.4117	0.5434	0.9371	0.7085	0.7068	0.9359	0.5330	0.2838	-0.2589
REG MATH	0.4667	0.3574	0.3682	0.4845	0.9207	0.6856	0.6849	0.9208	0.4771	0.2756	-0.2926
DATA INT	0.3929	0.3664	0.3602	0.4486	0.7514	0.5753	0.5716	0.7513	0.4463	0.1899	-0.2765
ANALREAS	0.5401	0.3886	0.3847	0.5192	0.7330	0.9718	0.9707	0.7339	0.5094	0.3250	-0.0553
LOGLREAS	0.6103	0.5644	0.5415	0.6674	0.5251	0.7014	0.6989	0.5245	0.6617	0.3216	-0.1716
SENTCOMP	0.6531	0.6467	0.6510	0.8447	0.5115	0.5256	0.5220	0.5125	0.8311	0.2995	-0.1023
READCOMP	1.0000	0.6056	0.6152	0.8605	0.5474	0.6163	0.6181	0.5478	0.8559	0.3088	-0.0882
ANALOGY	0.6056	1.0000	0.6637	0.8200	0.4261	0.4778	0.4756	0.4250	0.8200	0.2410	-0.1212
ANTONYM	0.6152	0.6637	1.0000	0.8796	0.4306	0.4683	0.4659	0.4308	0.8841	0.2038	-0.1417
VRB RGHT	0.8605	0.8200	0.8796	1.0000	0.5625	0.6147	0.6130	0.5627	0.9967	0.3046	-0.1330
QNT RGHT	0.5474	0.4261	0.4306	0.5625	1.0000	0.7532	0.7513	0.9994	0.5537	0.2942	-0.3050
AN RGHT	0.6163	0.4778	0.4683	0.6147	0.7532	1.0000	0.9984	0.7537	0.6050	0.3578	-0.0931
AN CONV	0.6181	0.4756	0.4659	0.6130	0.7513	0.9984	1.0000	0.7517	0.6029	0.3589	-0.0959
QNT CONV	0.5478	0.4250	0.4308	0.5627	0.9994	0.7537	0.7517	1.0000	0.5540	0.2917	-0.3035
VRB CONV	0.8559	0.8200	0.8841	0.9967	0.5537	0.6050	0.6029	0.5540	1.0000	0.2939	-0.1324
GPA	0.3088	0.2410	0.2038	0.3046	0.2942	0.3578	0.3589	0.2917	0.2939	1.0000	-0.0331
SEX	-0.0882	-0.1212	-0.1417	-0.1330	-0.3050	-0.0931	-0.0959	-0.3035	-0.1324	-0.0331	1.0000

BEST COPY AVAILABLE

CORRELATIONS CORRECTED FOR ATTENUATION FORM 3A

	AR3	LR3	CV3	AX3	NLR3	PI3	QUANT	REG MATH	DATA INT	ANALREAS	LOGLREAS	SENTCOMP
AR3	1.00	0.80	0.58	0.57	0.71	0.69	0.82	0.78	0.75	0.85	0.72	0.57
LR3	0.80	1.00	0.84	0.80	0.95	0.64	0.76	0.69	0.69	0.65	0.92	0.80
CV3	0.58	0.84	1.00	0.84	0.88	0.52	0.61	0.56	0.55	0.60	0.89	0.91
AX3	0.57	0.80	0.84	1.00	0.76	0.57	0.53	0.46	0.39	0.50	0.75	0.81
NLR3	0.71	0.95	0.88	0.76	1.00	0.70	0.75	0.67	0.77	0.65	0.84	0.74
PI3	0.69	0.64	0.52	0.57	0.70	1.00	0.64	0.63	0.55	0.64	0.63	0.52
QUANT	0.82	0.76	0.61	0.53	0.75	0.64	1.00	0.94	0.81	0.80	0.68	0.62
REG MATH	0.78	0.69	0.56	0.46	0.67	0.63	0.94	1.00	0.91	0.80	0.65	0.61
DATA INT	0.75	0.69	0.55	0.39	0.77	0.55	0.81	0.91	1.00	0.74	0.71	0.64
ANALREAS	0.85	0.65	0.60	0.50	0.65	0.64	0.80	0.80	0.74	1.00	0.69	0.56
LOGLREAS	0.72	0.92	0.89	0.75	0.84	0.63	0.68	0.65	0.71	0.69	1.00	0.85
SENTCOMP	0.57	0.80	0.91	0.81	0.74	0.52	0.62	0.61	0.64	0.56	0.85	1.00
READCOMP	0.72	0.85	0.89	0.79	0.84	0.59	0.69	0.60	0.57	0.66	0.88	0.89
ANALOGY	0.64	0.84	0.96	0.81	0.75	0.45	0.58	0.52	0.60	0.53	0.92	1.00
ANTONYM	0.52	0.67	0.80	0.60	0.56	0.34	0.49	0.45	0.50	0.45	0.74	0.85
VRB RGHT	0.64	0.82	0.93	0.77	0.75	0.50	0.63	0.57	0.60	0.58	0.88	1.06
QNT RGHT	0.82	0.74	0.60	0.50	0.75	0.64	1.08	1.09	1.00	0.82	0.70	0.64
AN RGHT	0.87	0.76	0.70	0.59	0.73	0.68	0.82	0.81	0.77	1.09	0.94	0.66
AN CONV	0.82	0.71	0.66	0.55	0.69	0.64	0.78	0.77	0.73	1.03	0.88	0.62
QNT CONV	0.78	0.71	0.58	0.47	0.71	0.61	1.03	1.03	0.95	0.78	0.66	0.61
VRB CONV	0.61	0.77	0.88	0.72	0.71	0.47	0.59	0.54	0.57	0.54	0.83	0.99
GPA	0.28	0.32	0.34	0.38	0.36	0.32	0.31	0.31	0.24	0.35	0.41	0.36
SEX	-0.13	-0.26	-0.16	-0.05	-0.20	-0.13	-0.28	-0.33	-0.35	-0.06	-0.22	-0.12

CORRELATIONS CORRECTED FOR ATTENUATION FORM 3A

	READCOMP	ANALOGY	ANTONYM	VRB RGHT	QNT RGHT	AN RGHT	AN CONV	QNT CONV	VRB CONV	GPA	SEX
AR3	0.72	0.64	0.52	0.64	0.82	0.87	0.82	0.78	0.61	0.28	-0.13
LR3	0.85	0.84	0.67	0.82	0.74	0.76	0.71	0.71	0.77	0.32	-0.26
CV3	0.89	0.96	0.80	0.93	0.60	0.70	0.66	0.58	0.88	0.34	-0.16
AX3	0.79	0.81	0.60	0.77	0.50	0.59	0.55	0.47	0.72	0.38	-0.05
NLR3	0.84	0.75	0.56	0.75	0.75	0.73	0.69	0.71	0.71	0.36	-0.20
PI3	0.59	0.45	0.34	0.50	0.64	0.68	0.64	0.61	0.47	0.32	-0.13
QUANT	0.69	0.58	0.49	0.63	1.08	0.82	0.78	1.03	0.59	0.31	-0.28
REG MATH	0.60	0.52	0.45	0.57	1.09	0.81	0.77	1.03	0.54	0.31	-0.33
DATA INT	0.57	0.60	0.50	0.60	1.00	0.77	0.73	0.95	0.57	0.24	-0.35
ANALREAS	0.66	0.53	0.45	0.58	0.82	1.09	1.03	0.78	0.54	0.35	-0.06
LOGLREAS	0.88	0.92	0.74	0.88	0.70	0.94	0.88	0.66	0.83	0.41	-0.22
SENTCOMP	0.89	1.00	0.85	1.06	0.64	0.66	0.62	0.61	0.99	0.36	-0.12
READCOMP	1.00	0.90	0.77	1.03	0.66	0.75	0.71	0.63	0.98	0.35	-0.10
ANALOGY	0.90	1.00	0.94	1.11	0.58	0.66	0.62	0.55	1.06	0.31	-0.16
ANTONYM	0.77	0.94	1.00	1.01	0.49	0.54	0.51	0.47	0.97	0.22	-0.15
VRB RGHT	1.03	1.11	1.01	1.00	0.62	0.68	0.64	0.59	1.05	0.32	-0.14
QNT RGHT	0.66	0.58	0.49	0.62	1.00	0.84	0.79	1.05	0.58	0.31	-0.32
AN RGHT	0.75	0.66	0.54	0.68	0.84	1.00	1.06	0.80	0.64	0.38	-0.10
AN CONV	0.71	0.62	0.51	0.64	0.79	1.06	1.00	0.75	0.60	0.36	-0.10
QNT CONV	0.63	0.55	0.47	0.59	1.05	0.80	0.75	1.00	0.55	0.29	-0.30
VRB CONV	0.98	1.06	0.97	1.05	0.58	0.64	0.60	0.55	1.00	0.29	-0.13
GPA	0.35	0.31	0.22	0.32	0.31	0.38	0.36	0.29	0.29	1.00	-0.03
SEX	-0.10	-0.16	-0.15	-0.14	-0.32	-0.10	-0.10	-0.30	-0.13	-0.03	1.00

54020-06557 • S81M.7 • 298024